

ATAS DO XXIII CONGRESSO

Da Sociedade Portuguesa de Estatística



Editores:

Maria de Fátima Salgueiro
Paula Vicente
Teresa Calapez
Catarina Marques
Maria Eduarda Silva

ATAS DO XXIII CONGRESSO DA SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Lisboa, 18 a 21 de outubro de 2017

Editores

Maria de Fátima Salgueiro

Paula Vicente

Teresa Calapez

Catarina Marques

Maria Eduarda Silva

Janeiro, 2020

Edições SPE

© 2020, Sociedade Portuguesa de Estatística

Editores: Maria de Fátima Salgueiro, Paula Vicente, Teresa Calapez,
Catarina Marques e Maria Eduarda Silva

Título: Atas do XXIII Congresso da Sociedade Portuguesa de Estatística

Editora: Sociedade Portuguesa de Estatística

Conceção Gráfica da Capa: Andreia Garcia (Iscte - Instituto Universitário
de Lisboa)

ISBN: 978-972-8890-46-9

Prefácio

Este é o Livro de Atas do XXIII Congresso da Sociedade Portuguesa de Estatística (SPE), que se realizou em Lisboa entre 18 e 21 de Outubro de 2017, nas instalações do ISCTE-Instituto Universitário de Lisboa.

Lisboa foi desta feita escolhida pela Sociedade Portuguesa de Estatística (SPE) para acolher o seu Congresso de 2017.

Lisboa, Janeiro de 2020

Os Editores

Agradecimentos

Aos seguintes colegas, pelo generoso trabalho de revisão dos artigos submetidos a este Livro de Atas, que em muito valorizou o conteúdo desta publicação:

- **Ana Paula Amorim**, Universidade do Minho
- **Ana Sousa Ferreira**, Universidade de Lisboa
- **Antónia Turkman**, Universidade de Lisboa
- **Carlos Tenreiro**, Universidade de Coimbra
- **Cláudia Silvestre**, Instituto Politécnico de Lisboa
- **Conceição Amado**, IST, Universidade de Lisboa
- **Cristina Miranda**, Universidade de Aveiro
- **Esmeralda Gonçalves**, Universidade de Coimbra
- **Graça Trindade**, Iscte - Instituto Universitário de Lisboa
- **Helena Ferreira**, Universidade da Beira Interior
- **Helena Mourinho**, Universidade de Lisboa
- **Isabel Alves Rodrigues**, IST, Universidade de Lisboa
- **Isabel Barão**, Universidade de Lisboa
- **Isabel Pereira**, Universidade de Aveiro
- **Isabel Silva Magalhães**, Universidade do Porto
- **Joana Leite**, Instituto Politécnico de Coimbra
- **José Dias Curto**, Iscte - Instituto Universitário de Lisboa
- **José Manuel G.Dias**, Iscte - Instituto Universitário de Lisboa

- **Lisete Sousa**, Universidade de Lisboa
- **Luís Antunes**, Universidade do Porto
- **Luís Machado**, Universidade do Minho
- **Manuel Scotto**, IST, Universidade de Lisboa
- **Manuela Neves**, ISA, Universidade de Lisboa
- **Margarida Cardoso**, Iscte - Instituto Universitário de Lisboa
- **Maria Almeida Silva**, Universidade de Lisboa
- **Maria da Graça Temido**, Universidade de Coimbra
- **Maria do Carmo Botelho**, Iscte - Instituto Universitário de Lisboa
- **Helena Carvalho**, Iscte - Instituto Universitário de Lisboa
- **Marília Antunes**, Universidade de Lisboa
- **Miguel Pereira**, Imperial College, London
- **Nazaré Mendes-Lopes**, Universidade de Coimbra
- **Paula Milheiro-Oliveira**, Universidade do Porto
- **Rui Menezes**, Iscte - Instituto Universitário de Lisboa
- **Sandra Dias**, Universidade de Trás-os-Montes e Alto Douro
- **Sebestyan Szabolcs**, Iscte - Instituto Universitário de Lisboa
- **Sofia Azevedo**, Faculdade de Ciências, Universidade de Lisboa

Um **agradecimento especial** é também devido aos colegas da **Direção da Sociedade Portuguesa de Estatística** que colaboraram diretamente na realização deste congresso e aos colegas das Comissões Científica e Organizadora do XXIII Congresso da Sociedade Portuguesa de Estatística.

Comissão Científica

- **Maria Eduarda Silva**, *Presidente da Sociedade Portuguesa de Estatística*, Faculdade de Economia, Universidade do Porto
- **Maria de Fátima Salgueiro**, Iscte - Instituto Universitário de Lisboa
- **Nazaré Mendes-Lopes**, Universidade de Coimbra
- **Conceição Amado**, Instituto Superior Técnico
- **Paulo M.M. Rodrigues**, Nova School of Business and Economics
- **José Manuel G. Dias**, Iscte - Instituto Universitário de Lisboa

Comissão Organizadora

- **Maria de Fátima Salgueiro**
- **Paula Vicente**
- **Teresa Calapez**
- **Catarina Marques**
- **Elizabeth Reis**

*Iscte - Instituto Universitário de Lisboa
e Business Research Unit (BRU - Iscte)*

Agradecimentos

Agradecemos às seguintes entidades o valioso apoio concedido para a realização do XXIII Congresso da SPE

- Banco de Portugal
- Edições Sílabo
- EPAL - Grupo Águas de Portugal
- Escolar Editora
- Fundação para a Ciência e a Tecnologia
- Instituto Nacional de Estatística
- Iscte - Executive Education
- Iscte - Instituto Universitário de Lisboa
- Produtos e Serviços de Estatística, PSE
- Sociedade Portuguesa de Estatística
- Turismo de Lisboa

Índice

Comparing Cox regression, parametric and flexible parametric models in the study of time to non-persistence in a chronic disease treatment	1
<i>Ana Rita Godinho, Cristina Rocha e Zilda Mendes</i>	
Avaliação de resultados em classificação supervisionada	13
<i>Ana Sousa Ferreira e Anabela Marques</i>	
Comparação bayesiana de testes de diagnóstico com dados densamente omissos ao acaso	31
<i>Carlos Daniel Paulino e Giovani L. Silva</i>	
O critério <i>Minimum Message Lenght</i> na estimação de modelos de mistura sobre dados mistos	45
<i>Cláudia Silvestre, Margarida G.M.S. Cardoso e Mário A.T. Figueiredo</i>	
Método das maiores observações anuais: Aplicação ao triplo-salto masculino	59
<i>Domingos Silva, Frederico Caeiro e Manuela Oliveira</i>	
Taxas de erros de tipos I e II de procedimentos não paramétricos alternativos à ANOVA com dois fatores para dados discretos	75
<i>Dulce G. Pereira e Anabela Afonso</i>	
Uma nova abordagem na avaliação da interacção genótipo × ambiente em espécies lenhosas de propagação vegetativa: o caso de clones de videira	89
<i>Elsa Gonçalves e Antero Martins</i>	
Propriedade de Taylor e curtose em modelos MA	105
<i>Esmeralda Gonçalves, Cristina Martins e Nazaré Mendes-Lopes</i>	

Números de clientes servidos e bloqueados em períodos de ocupação contínua de filas $M/M/1/n$ com bloqueio	117
<i>Fátima Ferreira, António Pacheco e Helena Ribeiro</i>	
Generalização do estimador de Hill, baseada na média de Lehmer: Resultados adicionais	129
<i>Ivanilda Cabral, Frederico Caeiro e M. Ivette Gomes</i>	
Modelos de sobrevivência aplicados à análise de acontecimentos múltiplos	145
<i>Ivo Sousa-Ferreira, Ana Maria Abreu e Cristina Simões Rocha</i>	
Modelagem de capturas em peso inflacionadas de zeros no Baixo Rio Amazonas	161
<i>Júlio César Pereira, Giovani L. Silva e Victória Isaac</i>	
Optimal re-sampled efficient frontier and examples	175
<i>Marcus Huber Mendes, Reinaldo Castro Sousa e Marco Aurélio Sanfins</i>	
Modelling (and forecasting) extremes in time series: A naive approach	189
<i>M. Manuela Neves e Clara Cordeiro</i>	
A importância dos conceitos e das classificações nas Estatísticas da Educação	203
<i>Nuno Rodrigues, Joaquim Santos, Carlos Malaca e Luísa Canto e Castro Loura</i>	
Omissões e dimensão da amostra: Impacto sobre medidas de qualidade do ajustamento em modelos de análise fatorial confirmatória	221
<i>Paula C.R. Vicente e Maria de Fátima Salgueiro</i>	
Os sindicatos no feminino: Um ensaio sobre diferentes formas de visualização	235
<i>Paulo Marques Alves e Maria do Carmo Botelho</i>	

Distribuição limite conjunta da soma e do máximo de variáveis inteiras 249

Sandra Dias e Maria da Graça Temido

Estatísticas ordinais de uma amostra aleatória: O caso de dimensão de amostra com distribuição binomial negativa 263

Sandra Mendonça e Délia Gouveia-Reis

Autores 275

Comparing Cox regression, parametric and flexible parametric models in the study of time to non-persistence in a chronic disease treatment

Ana Rita Godinho

Centro de estudos e avaliação em saúde (CEFAR), Associação Nacional das Farmácias (ANF), *Ana.Godinho@anf.pt*

Cristina Rocha

Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal e Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal, *cmrocha@fc.ul.pt*

Zilda Mendes

Centro de estudos e avaliação em saúde (CEFAR), Associação Nacional das Farmácias (ANF), *Zilda.Mendes@anf.pt*

Keywords: Chronic Disease Treatment Persistence; Survival Analysis; Royston and Parmar Flexible Models; Parametric Models; Cox Model.

Abstract: The Cox model is the most frequently used procedure for the analysis of survival data, due to its not very restrictive assumptions. However, if found to be adequate, a parametric model would lead to more precise estimates of the regression parameters. Yet, sometimes, the parametric models may not be sufficiently flexible to adequately represent the baseline hazard function. Recently, a new class of flexible parametric models has become available. The study aims to compare the Cox model, traditional parametric models (Weibull and log-logistic) and flexible parametric models in the analysis of time to non-persistence in a chronic disease treatment.

1 Introduction

Due to population ageing, there has been an increase in the incidence of chronic diseases usually associated with debilitating or physically painful conditions [1], leading to a reduction in the patient's quality of life. Although developed countries have the oldest population profiles, less developed countries have a rapidly aging population. Thus, quality of life specially associated with chronic diseases, becomes increasingly a matter of public health [2]. Several studies show that medication persistence (i.e. the act of continuing treatment for the prescribed period) play a crucial role in improving health outcomes [3]. Therefore, it is essential to understand the factors that influence persistence in a treatment.

The Cox regression model [4] is the most frequently used method for analysing survival data. This model is semiparametric, since the underlying distribution of survival time is unspecified, which makes it so popular in medical sciences. On the other hand, if found to be adequate, a parametric regression model would lead to more precise estimates than the Cox model [5]. Nonetheless, even though the traditional parametric models offer advantages over the Cox model, they may not be sufficiently flexible to adequately represent the hazard function of each group of patients.

The flexible parametric models, proposed by Royston and Parmar [6, 7], are generalizations of the traditional parametric models, which introduce more flexibility in the form of the survival distribution they can model. The aim of this study is to use the Cox model, two parametric models (Weibull and log-logistic) and flexible parametric models (proportional hazards and proportional odds) to evaluate the effect of the patient's age, whether the patient lives alone or not and the type of treatment to which the patient is subjected, on the time to non-persistence in a chronic disease treatment. Another goal is to identify the model that best describes the hazard function associated with each group of patients and, thus, produces more precise estimates of the adjusted hazard ratios.

2 Methods

2.1 Data source

Data were obtained from an observational prospective cohort study, involving 360 individuals with a specific chronic disease. The patients were recruited in the Portuguese community pharmacies where their medication was purchased and were followed for a maximum period of 18 months (from January 2011). The event of interest was the non-persistence in the treatment of a chronic disease, i.e., the untimely discontinuation of the treatment. Therefore, the response variable was the time to non-persistence, defined as the time (in days) from initiation to discontinuation of the treatment. During recruitment, several sociodemographic and health-related variables were collected. According to the method for variable selection proposed by Collett [8], the only variables that showed significant influence on time to non-persistence were: Age, Living alone and Treatment (monthly or weekly treatment), and therefore, those were the only ones that were included in the analysis.

2.2 Cox proportional hazards model

The Cox PH model is the most widely used procedure in survival analysis. In this model, the hazard function for a vector of covariates, $\mathbf{x} = (x_1, \dots, x_p)$, is:

$$h(t; \mathbf{x}) = h_0(t)e^{\beta' \mathbf{x}}$$

where $\beta = (\beta_1, \dots, \beta_p)$ is the vector of regression coefficients.

This model is semiparametric, which means that the baseline hazard function, $h_0(t)$, is not specified. Through the hazard ratio (HR), it is possible to compare two individuals with covariate patterns \mathbf{x}_1 and \mathbf{x}_2 , for which only the value of one covariate differs, x_j :

$$\frac{h(t; \mathbf{x}_1)}{h(t; \mathbf{x}_2)} = \exp(\beta_j(x_{1j} - x_{2j}))$$

One constraint of the Cox PH model is its proportional hazards assumption. It means that the hazard ratio between two individuals with different vectors of covariates is constant over time, so, if this assumption is violated, the results may not be reliable.

2.3 Parametric models

Unlike the Cox PH model, in parametric models the response variable (survival time) is assumed to follow a distribution with unknown parameters that are estimated from the data. When there are strong indications that a certain distribution is appropriate, it is preferable to use these models, as they are more efficient and yield results more consistent with the theoretical survival curve. In addition, not all parametric models satisfy a proportional hazards assumption, instead many parametric models are accelerated failure time (AFT) models or even proportional odds (PO) models.

In the AFT models, the covariates have a multiplicative effect on the survival time. In this case, the survival function for a vector of covariates, \mathbf{x} , is expressed as follows:

$$S(t; \mathbf{x}) = S_0(t \exp(\boldsymbol{\alpha}' \mathbf{x}))$$

On the other hand, the PO models satisfy a proportional odds assumption and the covariates have a multiplicative effect on the survival odds. The odds of survival beyond time t for a vector of covariates \mathbf{x} is given by:

$$\frac{S(t; \mathbf{x})}{1 - S(t; \mathbf{x})} = e^{\eta} \frac{S_0(t)}{1 - S_0(t)}$$

where $\eta = \boldsymbol{\beta}' \mathbf{x}$.

2.4 Royston and Parmar flexible models

In this paper, we compare Cox PH and traditional parametric (Weibull and log-logistic) models with an alternative class of models proposed by Royston and Parmar, the flexible parametric models [6, 7]. To

obtain the flexible parametric models the approach taken by the authors is to model a transformation of the survival function as a natural cubic spline function of log time:

$$g[S(t; \mathbf{x})] = s(\ln(t), \gamma) + \beta' \mathbf{x}$$

The natural cubic spline is constrained to be linear beyond its boundary knots k_{min} , k_{max} and can have m internal knots k_1, \dots, k_m (with $k_1 > k_{min}$ and $k_m < k_{max}$). A natural cubic spline for z is given by:

$$s(z, \gamma) = \gamma_0 + \gamma_1 z + \gamma_2 v_1(z) + \dots + \gamma_{m+1} v_m(z)$$

where $v_j(z) = (z - k_j)_+^3 - \lambda_j(z - k_{min})_+^3 - (1 - \lambda_j)(z - k_{max})_+^3$ and $\lambda_j = (k_{max} - k_j)/(k_{max} - k_{min})$ and $(z - a)_+ = \max(0, z - a)$.

In this study, we focus in the flexible parametric models with proportional hazards (PH) and proportional odds assumptions (PO), which are generalizations of the Weibull and log-logistic models, respectively. To find the optimal number of internal knots, we used the Akaike information criterion (AIC). As for their location, we selected the centile-based positions, as suggested by the authors, i.e., the centiles of the distribution of the uncensored log-survival times.

All statistical analysis was performed using R statistical software v3.0.1. The flexsurv package was used for the flexible parametric analysis.

3 Results

A total number of 360 patients with a specific chronic disease were included in this study. Of the complete cohort, 80 patients (22.3%) lived by themselves and 242 patients (67.2%) were under a weekly treatment. About 36.9% of the patients were aged between 60 and 70 years and 27.8% were aged above 70 years. The patient characteristics are summarized in Table 1. At the end of the follow-up

period, the event of interest was observed for 275 (76.9%) patients. Of the 85 patients for whom the event was not observed, 6 were lost to follow-up, consequently 79 patients remained persistent until the end of the study.

Table 1: Patients characteristics

Variable	Category	No. of patients (%)
Age	≤ 60 years	127 (35.3%)
	60 to 70 years	133 (36.9%)
	> 70 years	100 (27.8%)
Lives alone	No	280 (77.7%)
	Yes	80 (22.3%)
Treatment	Monthly	118 (32.8%)
	Weekly	242 (67.2%)

3.1 Traditional parametric and Flexible parametric models

Table 2 compares the AIC value for the multivariable PH and PO flexible parametric models with up to four internal knots. The models with no internal knots ($m = 0$) are equivalents of the Weibull and the log-logistic model, in PH and PO modelling respectively. The increase in the number of internal knots, up to three, leads to a decrease in the AIC value associated with each model, the lowest AIC values were found under the PH flexible parametric models with two and three internal knots (respectively, $AIC = 3576.56$ and $AIC = 3574.67$). For these two models not only are the AIC values close but also the parameters estimates remain unchanged as shown in Table 3, thus, the inclusion of a third knot results in an unnecessary increase of the curve's complexity. On this account, the optimal number of internal knots was found to be $m = 2$ under the

PH flexible parametric model.

Table 2: AIC values for multivariable flexible parametric models

No. of knots	PH	PO
0	3653.99	3623.74
1	3593.00	3594.60
2	3576.56	3578.99
3	3574.67	3578.73
4	3576.06	3580.38

Table 3: Parameters estimates in the PH multivariable flexible parametric models

Variable	$m = 0$	$m = 1$	$m = 2$	$m = 3$
Age – 60 to 70 years	-0,382	-0,360	-0,364	-0,363
Age – >70 years	-0,013	-0,022	-0,015	-0,016
Lives alone	0,440	0,373	0,390	0,387
Treatment	0,262	0,261	0,260	0,259
γ_0	-4,784	-8,205	-11,189	-14,184
γ_1	0,820	1,788	2,738	3,778
γ_2	-	0,097	0,374	0,580
γ_3	-	-	-0,276	-0,250
γ_4	-	-	-	-0,067

The survival curves estimated with the univariable PH flexible parametric models with up to three internal knots are shown in Figure 1. The inclusion of one internal knot ($m = 1$) produces a clear change in the estimate of the survival curves associated with each level of

the three variables under analysis. For all groups, the curves become more flexible at the beginning of the follow-up and get closer to the Kaplan-Meier estimates. The addition of a second knot affects the curvature of the estimates which, although in a less evident way, improves its fit in relation to the models with only one knot, leading to an overlapping of this curves and the respective Kaplan-Meier estimate.

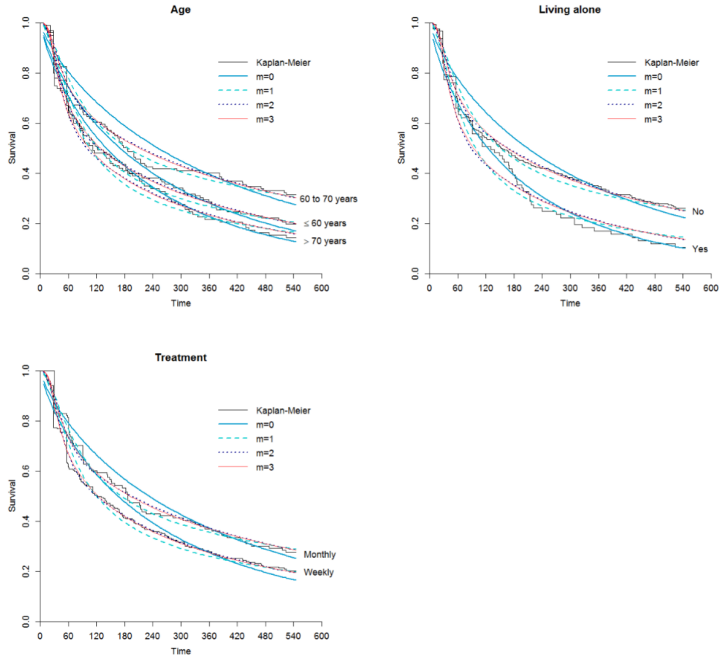


Figure 1: Survival curves: comparison between Kaplan-Meier estimates and the curves obtain from the PH flexible parametric models with m internal knots

In turn, the models with three internal knots do not change significantly the form of the survival functions estimates when compared to the previous models. This result is consistent with the earlier analysis of the AIC values in the multivariable analysis, suggesting the inclusion of the third knot to be unnecessary.

3.2 Cox and Flexible parametric models

The results of Cox PH model and PH flexible parametric models with two internal knots are shown in Table 4. In general, the hazard ratios are very similar for the two models.

Table 4: Multivariable analysis of Cox and PH flexible parametric model with 2 knots (RP model)

Variable	Category	Cox model	RP model
		HR[95% CI]	HR[95% CI]
Age	≤ 60 years	1	1
	60 to 70 years	0.697 [0.522; 0.930]	0.695 [0.521; 0.928]
	> 70 years	0.979 [0.722; 1.327]	0.985 [0.727; 1.335]
Lives alone	No	1	1
	Yes	1.474 [1.110; 1.957]	1.478 [1.113; 1.962]
Treatment	Monthly	1	1
	Weekly	1.331 [1.026; 1.727]	1.297 [1.000; 1.683]

According to the results, with the selected PH flexible parametric model we found that for patients with the same values in the remaining variables, a patient with 60 to 70 years has a smaller risk of becoming non-persistent (less 30.5%), than a patient with 60 years or younger (HR 0.695; 95% CI [0.521; 0.928]). A patient with more than 70 years has roughly the same risk as a patient with 60 years or younger (HR 0.985; 95% CI [0.727; 1.335]). Living alone increases the risk of non-persistence by 47.8% (HR 1.478;

95% CI [1.113; 1.962]). A patient under weekly treatment has a 29.7% greater risk of becoming non-persistent than a patient with a monthly treatment (HR 1.297; 95% CI [1.000; 1.683]).

4 Discussion

Our goal was to compare Cox, traditional parametric and flexible parametric models applied to the study of time to non-persistence in a chronic disease treatment. Upon finding the model that best described the data, we intended to evaluate the impact of the patient's age, the fact that the patient lives alone or not and the type of treatment on survival time, i.e., on time to non-persistence in the treatment. Researchers in the field of life sciences are usually more interested in the Cox proportional hazards model rather than parametric models. However, if the Cox model's assumptions do not hold, the model can lead to biased estimates, thus, this model might not be appropriate in some situations. Besides, when a certain distribution is found to be adequate, the corresponding parametric regression model provides more accurate estimates. As mentioned by Kleinbaum [5], although it may be preferable to use a parametric model, most of the time we are not sure which is the proper distribution and since the Cox model is robust, generating results very close to the adequate parametric model, it becomes greatly popular. Even when we are sure about which parametric model to use, it may not be flexible enough to describe the data adequately. In recent years, a new family of models was proposed and developed, the flexible parametric models. These models introduce a greater flexibility to the shape of the survival distribution. According to the results, the inclusion of internal knots improves the estimation process of the survival functions in both the PH and the PO flexible parametric models, when compared to the models without internal knots (which are equivalent to the Weibull and log-logistic, respectively). The increase in the number of knots, to a maximum of three, resulted in the decrease of the AIC value and in the increase of flexibility of

the estimated survival curves, which got closer to the Kaplan-Meier estimates. The PH flexible parametric models with two and three internal knots seem to be the most adequate and the ones that best describe the data. Since the parameters estimates and AIC values of both models are very similar, we opted for the most parsimonious model, which is the PH flexible parametric model with two internal knots. Thus, based on this model, for individuals with the same value in the remaining variables, it is estimated that a patient aged between 60 and 70 years has a lower risk and a patient with more than 70 years has around the same risk of becoming non-persistent than a patient with 60 years or younger. When comparing patients in the same age group and under the same treatment, a patient who lives alone has a higher risk of discontinuing the treatment than a patient who lives with someone else. Lastly, it is estimated that a patient under a weekly treatment has a higher risk of becoming non-persistent than a patient under a monthly treatment, for patients in the same age group and living in the same conditions. Overall, the hazard ratios estimated with the selected flexible parametric model and with the Cox model are close, as expected. Nevertheless, using flexible modelling to analyse the effect of a set of covariates on time to non-persistence in the treatment of a chronic disease, leads to more precise estimates than by using other families of models. Thus, this methodology contributes for a better understanding of the phenomenon in study over time.

Acknowledgements

Ana Rita Godinho and Zilda Mendes's work was partially supported by CEFAR. Cristina Rocha's work was partially supported by FCT Portugal UID/MAT/00006/2013.

References

- [1] Ferreira, L.N., Ferreira, P.L., Pereira, L.N., et al. (2014). EQ-5D Portuguese population norms. *Quality of Life Research* 23, 425–430.
- [2] World Health Organization, National Institute of Health, National Institute on Aging, et al. (2011) *Global Health and Aging*. 11–7737.
- [3] Cramer, J.A., Roy, A., Burrell, A., et al. (2008). Medication Compliance and Persistence: Terminology and Definitions. *Value in Health* 11, 44–47.
- [4] Cox, D.R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B (Methodological)* 34(2), 187–220.
- [5] Kleinbaum, D.G., Klein, M. (2005). *Survival Analysis: A Self-Learning Text* 2nd ed. Springer, New York.
- [6] Royston, P., Parmar, M.K.B. (2002) Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 21, 2175–22197.
- [7] Royston, P., Lambert, P.C. (2011) *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model* 1st ed. Stata Press, Texas.
- [8] Collett, D. (2003) *Modelling Survival Data in Medical Research* 2nd ed. Boca Raton: Chapman and Hall/CRC.

Avaliação de resultados em classificação supervisionada

Ana Sousa Ferreira

Faculdade de Psicologia, Universidade de Lisboa, Business Research Unit (BRU-IUL), Lisboa, Portugal,
asferreira@psicologia.ulisboa.pt

Anabela Marques

Escola Superior de Tecnologia do Barreiro, IPS, CIIAS, Barreiro, Portugal,
anabela.marques@estbarreiro.ips.pt

Palavras-chave: Avaliação de resultados; Classificação Supervisionada; Combinação de modelos.

Resumo: Em problemas discretos de classificação supervisionada observa-se, frequentemente, que as observações mal classificadas são diferentes para diferentes modelos. Deste modo, a abordagem pela combinação de modelos tem vindo a ser considerada uma mais valia neste domínio. A avaliação de resultados em classificação baseia-se, habitualmente, na taxa de casos bem classificados. No entanto, alguns autores têm vindo a advertir que esta medida pode não analisar corretamente a qualidade de um modelo. Neste trabalho, pretendemos explorar a avaliação de desempenho de novos modelos combinados, comparando a medida de avaliação mais usual com outros tipos de medidas como a Sensibilidade, Especificidade ou Precisão, Medidas de associação ou concordância ou o Índice de Huberty.

1 Introdução

Em Estatística, fala-se de um problema de classificação supervisionada quando se pretende identificar qual a classe, entre várias definidas *a priori*, a que pertence uma nova observação, baseando-se na

informação fornecida por uma amostra, onde a classe de cada observação é conhecida. Por exemplo, quando se pretende atribuir um diagnóstico a um certo paciente, descrito por um conjunto de características observadas (sexo, pressão arterial, presença ou ausência de alguns sintomas, . . .), entre meningite viral ou bacteriana ou quando se precisa de decidir se um dado *email* pertence à classe de *emails* “spam” ou “não spam”. Em qualquer dos exemplos referidos, para identificar a classe a que pertence a nova observação, utiliza-se a informação de uma amostra, denominada habitualmente amostra de treino, tentando perceber se o “perfil” da nova observação, será mais provável de ocorrer na Classe 1 ou na Classe 2.

No caso discreto, os resultados que podem ser observados são denominados por estados. Exemplificando, no caso mais simples de apenas duas variáveis binárias (0 - ausência do sintoma e 1 - presença do sintoma) podem ocorrer os estados seguintes: 00, 01, 10 e 11. Então, os resultados observados numa amostra de treino podem ser apresentados como na Tabela 1:

Tabela 1: Exemplo de estados observados numa amostra de treino

	Estados	Classe 1	Classe 2
1	00	4	0
2	01	5	1
3	10	0	4
4	11	1	5
	Total	10	10

No caso discreto, o modelo mais natural é o Modelo Multinomial Completo (MMC) onde a probabilidade de ocorrer um certo estado se a observação pertencer a uma determinada classe é estimada pela frequência relativa observada na amostra-treino, em cada classe ([5]). Contudo, quando o número de variáveis consideradas aumenta um pouco, o número de estados possíveis sofre, de imediato, um enorme incremento. Note-se, por exemplo que, no caso

mais simples de variáveis binárias, se forem consideradas 10 variáveis, teremos $2^p = 2^{10} = 1024$ estados possíveis, exigindo amostras de grandes dimensões para permitir a estimação de todos os parâmetros do modelo.

Deste modo, em classificação supervisionada, no caso discreto, existe frequentemente um problema de dimensionalidade, denominado mesmo na literatura como “a maldição da dimensionalidade”:

- Na generalidade dos modelos, o número de parâmetros a ser estimado é demasiado grande;
- Em Ciências Sociais e Humanas, onde o caso discreto tem grande prevalência, não raramente as amostras têm pequena dimensão.

Consequentemente, gera-se facilmente um número elevado de estados não observados, dificultando a estimação de todos os parâmetros. Este problema conduz a que a maior parte dos métodos revelem um fraco desempenho, especialmente quando as classes são pouco separadas e não balanceadas ([6]). Deste modo, em problemas de classificação discretos, a abordagem pela combinação de modelos tem vindo a ser referida como uma *mais-valia*, resultante de os erros de má classificação observados em diferentes modelos tenderem a ocorrer em objetos diferentes ([2], [7], [11]).

Quando se compara o desempenho destes novos modelos combinados com os modelos originais, usa-se geralmente a Taxa de casos bem classificados ou de casos mal classificados. No entanto, esta medida de avaliação pode não analisar corretamente a qualidade de um modelo, particularmente quando as classes são não balanceadas. Neste trabalho, pretendemos explorar a avaliação de resultados em classificação supervisionada, comparando a taxa de casos bem classificados com outros tipos de medidas ([4], [9]).

2 Combinação de modelos

Geralmente, em face de um problema de classificação complexo, estimam-se diversos modelos e, posteriormente, um único modelo é selecionado, baseado num determinado critério de validação. Contudo, os modelos descartados contêm frequentemente alguma informação importante sobre o problema de classificação, que se perde pelo facto de se considerar um único modelo ([2]). Por outro lado, verifica-se muitas vezes que as observações mal classificadas são diferentes para diferentes modelos. Este conhecimento tem conduzido a um número crescente de publicações sobre abordagens de combinação de modelos, ainda que referenciadas sob nomes diversos como *Blending*, *Bagging* e *Arcing* entre outros ([8]).

Em problemas de classificação discretos, consideram-se dois modelos de referência: o já referido Modelo Multinomial Completo (MMC) e o Modelo de Independência Condicional de ordem um (MIC) que considera as variáveis independentes dentro de cada classe, reduzindo assim o número de parâmetros a estimar de $2^p - 1$ para p , em cada classe.

Na abordagem de combinação de modelos proposta por Sousa Ferreira ([11]) e continuada por Marques ([7]) consideraram-se combinações lineares de dois modelos de referência no campo discreto. Inicialmente, Sousa Ferreira ([11]) propôs uma combinação linear entre os modelos de referência acima mencionados, MMC e MIC. Esperava-se, naturalmente, que esses dois modelos conduzissem a classificadores diferentes em muitas circunstâncias, dado que o primeiro pressupõe a existência de relações entre as p variáveis binárias e o segundo, considera que dentro de cada classe as p variáveis são independentes. O modelo combinado MMC-MIC resulta da combinação linear entre os dois modelos usando um único coeficiente β , $0 \leq \beta \leq 1$, conduzindo a um modelo intermédio entre MMC e MIC. As várias estratégias adoptadas para estimar β produzem diferentes modelos combinados ([2]). Num segundo momento, verificando que o modelo MMC revela grande dificuldade em estimar todos os parâmetros do modelo quando as amostras têm pequena dimensão,

Marques ([7]) desenvolveu uma combinação linear entre o Modelo Gráfico Decomponível (MGD) ([3]) e o modelo MIC, usando também um único coeficiente β , com valores no intervalo $[0,1]$. O modelo MGD considera as interações mais importantes entre pares de variáveis para estimar a função de probabilidade por classe, utilizando uma estrutura de árvore (grafo), que se baseia na informação mútua. O algoritmo considerado foi o proposto por Chow e Liu ([3]). Também neste caso se esperava que estes dois modelos conduzissem a classificadores diferentes, uma vez que o primeiro pressupõe a existência de interações entre as p variáveis binárias e o segundo, considera que dentro de cada classe as p variáveis são independentes. No caso de múltiplas classes *a priori*, ambos os modelos combinados, MMC-MIC e MGD-MIC, consideram o Modelo de Emparelhamento Hierárquico (MHIERM) que decompõe um problema de múltiplas classes em múltiplos problemas de duas classes ([2], [11]).

A abordagem de combinação de modelos proposta por Sousa Ferreira e continuada por Marques foi avaliada comparativamente com outros algoritmos existentes quer sobre dados reais quer simulados ([7],[8],[11]) revelando uma boa capacidade preditiva em casos de amostras de pequena ou moderada dimensão.

Neste trabalho, pretendemos continuar a explorar os resultados desta abordagem de combinação de modelos, usando outras medidas de avaliação da qualidade dos modelos ([4], [9]).

3 Medidas de avaliação

Na literatura de Estatística, a avaliação do desempenho de qualquer modelo de classificação supervisionada baseia-se, genericamente, na diagonal da matriz de confusão que confronta as classes preditas pelo modelo com as classes originais.

Diversas medidas de desempenho de um modelo podem ser definidas a partir dessa matriz, sendo tradicionalmente usadas a Taxa de casos bem classificados ou de casos mal classificados, estimadas por substituição, amostra-teste ou validação cruzada. Diferentes auto-

res têm vindo a referir, contudo, que estas estatísticas tradicionais de avaliação de resultados em classificação podem não analisar corretamente a qualidade de um algoritmo ou modelo ([4], [9], [10]).

Num problema de classificação discreto, com duas classes, tem-se a matriz de confusão apresentada na Tabela 2:

Tabela 2: Matriz de Confusão

	Classes preditas		
		1	2
Classes verdadeiras	1	a	b
	2	c	d

onde:

a - nº de casos bem classificados na classe 1

b - nº de casos da classe 1 classificados na classe 2

c - nº de casos da classe 2 classificados na classe 1

d - nº de casos bem classificados na classe 2

Em Medicina, os valores de a , b , c e d são denominados habitualmente por *Verdadeiros Positivos*, *Falsos Negativos*, *Falsos Positivos* e *Verdadeiros Negativos*, respetivamente. Esta terminologia, que se generalizou a muitos outros campos de aplicação, deriva de, por exemplo, se constatar que um exame complementar de diagnóstico indica que um certo sujeito está doente mas, na realidade, o sujeito está saudável. Teremos, então, um caso de *Falso Positivo*. Algumas medidas de avaliação em classificação estão associadas a este tipo de problemas de classificação.

Na Tabela 3 apresentam-se algumas medidas de avaliação baseadas na matriz de confusão. A *Taxa de casos bem classificados* ou *Acuracia* (Ac) é a medida mais comumente usada e mede a eficiência global do modelo. Na verdade, a *Acuracia* pretende responder à questão: “Globalmente, com que frequência o modelo de classificação decide corretamente?”

Tabela 3: Medidas de avaliação baseadas na matriz de confusão

Medidas	Definição
<i>Taxa de casos bem classificados</i> ou <i>Acuracia</i>	$\frac{a+d}{a+b+c+d}$
<i>Taxa de casos bem classificados na classe 1</i> ou <i>Sensibilidade</i>	$\frac{a}{a+b}$
<i>Taxa de casos bem classificados na classe 2</i> ou <i>Especificidade</i>	$\frac{d}{c+d}$
<i>Precisão</i>	$\frac{a}{a+c}$

A *Taxa de casos bem classificados na classe 1* é também denominada por *Sensibilidade* e mede a eficiência na classe 1 e a *Taxa de casos bem classificados na classe 2* é também denominada por *Especificidade* e mede a eficiência na classe 2. Como referido anteriormente, se um exame complementar de diagnóstico indicar que um certo sujeito está doente mas, na realidade, esse sujeito estiver saudável, temos um caso de *Falso Positivo*, pelo contrário, se esse sujeito estiver mesmo doente, temos um caso de *Verdadeiro Positivo*. Do mesmo modo, se o exame complementar de diagnóstico indicar que o sujeito não está doente e, de facto, esse sujeito estiver saudável, estamos perante um caso *Verdadeiro Negativo*. Naturalmente, um bom modelo de classificação deverá ser capaz de identificar quer os casos de *Verdadeiro Positivo* quer os de *Verdadeiro Negativo*.

A *Sensibilidade* é, exatamente, a taxa de casos *Verdadeiro Positivo* e a *Especificidade* a taxa de casos *Verdadeiro Negativo*, respondendo respetivamente às questões “Se um sujeito pertence à Classe 1, qual a frequência com que o modelo de classificação consegue identificar corretamente a classe desse sujeito?”, e, “Se um sujeito pertence à Classe 2, qual a frequência com que o modelo de classificação con-

segue identificar corretamente a classe desse sujeito?”. Finalmente, a *Precisão*, também denominada por valor preditivo positivo, mede a exatidão do modelo respondendo a outra questão: “Entre os casos que o modelo classificou como *Positivos*, isto é, pertencentes à Classe 1, quantos efetivamente o são?”. Um valor de *Precisão* elevado revela, pois, um modelo que é um bom preditor.

As medidas de avaliação usadas, em geral, não fornecem um equilíbrio entre os casos falsos positivos (*c*) e os falsos negativos (*b*). As medidas de avaliação combinadas, apresentadas na Tabela 4, tentam obter uma melhor paridade entre eles.

Tabela 4: Medidas de avaliação combinadas

Medidas	Definição
<i>Taxa de casos bem classificados balanceada</i>	$\frac{\text{Sensibilidade} + \text{Especificidade}}{2}$
<i>Média Geométrica entre Sensibilidade e Especificidade</i>	$\sqrt{\text{Sensibilidade} \times \text{Especificidade}}$
<i>Medida F</i>	$\frac{2 \times \text{Sensibilidade} \times \text{Precisão}}{\text{Sensibilidade} + \text{Precisão}}$

A *Taxa de casos bem classificados balanceada* ou *Acuracia balanceada* é a média aritmética entre a *Sensibilidade* e a *Especificidade* e, comparada com a *Acuracia global*, tenderá a ser menor quando o modelo não consegue classificar igualmente bem as duas classes. A *Média Geométrica* entre as duas medidas mede o equilíbrio entre a classificação nas duas classes. Um valor de *Média Geométrica* baixo indica um desempenho fraco na classe dita positiva (geralmente, considerada como classe de maior interesse). A *Medida F* combina as medidas *Sensibilidade* e *Precisão*, mesmo quando as classes de da-

dos são verdadeiramente desequilibradas. As medidas de avaliação já apresentadas anteriormente, sendo genericamente taxas, simples ou combinadas, variam naturalmente no intervalo $[0,1]$.

Um outro tipo de medidas de avaliação, que indicam a associação ou o acordo entre classes verdadeiras e preditas têm vindo a ser referidas por alguns autores. Por outro lado, parece ser relevante avaliar a melhoria efetiva que um modelo introduz relativamente à regra da maioria. Estas medidas de avaliação menos tradicionais em classificação supervisionada são apresentadas na Tabela 5.

Tabela 5: Outro tipo de medidas de avaliação

Medidas	Definição
<i>Coefficiente ϕ</i>	$\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$
<i>Estatística K de Cohen</i>	$\frac{A_c - P_{acaso}}{1 - P_{acaso}}, \text{ onde}$ $P_{acaso} = \left(\frac{a+b}{N} \times \frac{a+c}{N}\right) + \left(\frac{c+d}{N} \times \frac{b+d}{N}\right)$ $\text{e } N = a + b + c + d$
<i>Índice de Huberty</i>	$\frac{P_{cc} - P_m}{1 - P_m}, \text{ onde}$ $P_{cc} - \% \text{ casos corretamente classificados e}$ $P_m - \% \text{ casos corretamente classificados de acordo com a regra da maioria}$

O *Coefficiente ϕ* é uma conhecida medida de associação entre duas variáveis binárias, podendo tomar valores no intervalo $[-1,1]$. O sinal positivo deste coeficiente indica um maior número de casos em que o modelo de classificação decidiu corretamente e, o sinal negativo, pelo contrário, revela que existem mais casos de decisão incorreta. Por outro lado, a *Estatística K de Cohen* pode ser definida como

a proporção de acordo entre duas classificações após ser retirada a proporção de acordo devida ao acaso, podendo também tomar valores no intervalo $[-1,1]$. Por último, o *Índice de Huberty* avalia o desempenho de um modelo como o grau de correção da classificação realizada, comparando com a percentagem de casos bem classificados pela regra da maioria, sendo definido como a razão entre a melhoria efectiva e a melhoria possível na classificação. Este índice é a única medida de avaliação apresentada que pode tomar valores fora do intervalo $[-1,1]$.

4 Resultados Numéricos

Neste estudo, analisaram-se dados simulados, com duas classes e quatro variáveis binárias e consideraram-se três importantes fatores que influenciam o desempenho dos modelos: dados balanceados ou não balanceados, separabilidade das classes (baixa ou elevada) e dimensão das amostras (pequena ou grande). Considerando os oito cenários referidos, especificam-se seguidamente os valores considerados para cada fator: *i.* Equilíbrio - Classes balanceadas quando $n_1 = n_2$ e não balanceadas quando $n_1 = \frac{1}{9} \times n_2$; *ii.* Separabilidade - sendo medida pelo Coeficiente de Afinidade ([1]) definido no intervalo $[0,1]$. Este coeficiente mede a afinidade ou semelhança entre as classes pelo que, quanto mais pequeno for o seu valor, mais separadas são as classes consideradas e, por isso, a tarefa do modelo de classificação fica simplificada. Deste modo, considerou-se separabilidade baixa quando o coeficiente de afinidade toma valores superiores a 0,7 e elevada quando este coeficiente toma valores inferiores a 0,3; *iii.* Dimensão das amostras - pequena quando $n=60$ e grande quando $n=400$.

Considerando os dois graus de separabilidade das classes (Baixa ou Elevada), os dados em análise foram simulados segundo a Distribuição Multinomial com os parâmetros, isto é, as probabilidades de ocorrência das quatro variáveis preditoras binárias, apresentados na Tabela 6.

Tabela 6: Parâmetros da Distribuição Multinomial usados na simulação dos dados, de acordo com o grau de separabilidade (Baixa ou Elevada) entre as duas classes consideradas

Separab.	C_1	C_2
<i>Baixa</i>	(0,5;0,5;0,5;0,5;0,5;0,5;0,5;0,5)	(0,5;0,5;0,5;0,5;0,5;0,5;0,5;0,5)
<i>Elevada</i>	(0,1;0,9;0,7;0,3;0,2;0,8;0,6;0,4)	(0,9;0,1;0,3;0,7;0,8;0,2;0,4;0,6)

No estudo apresentado, para cada um dos oito cenários considerados, geraram-se 10 réplicas. Baseados nos 80 conjuntos de dados gerados, pretendemos averiguar a vantagem comparativa do modelo combinado MGD-MIC, usando diversas medidas de avaliação do desempenho. As medidas de avaliação de desempenho dos modelos foram todas estimadas por *2-fold cross validation*. O desempenho dos modelos, simples ou combinados, são apresentados nas Tabelas 7, 8, 9 e 10, onde se mostram os resultados médios intra-cenários (e respetivo desvio-padrão), destacando-se a negrito o melhor resultado obtido em cada cenário.

Na Tabela 7, podemos notar que, no caso de maior complexidade, quase todas as medidas elegem o modelo combinado como o melhor modelo, embora a sua capacidade preditiva seja apenas ligeiramente superior à dos modelos originais. Note-se, ainda, que neste caso de classes balanceadas, $\phi = Kappa = I.Huberty$. Quando a separabilidade é elevada, o modelo combinado revela um desempenho muito semelhante ao do modelo MIC, ambos demonstrando uma excelente capacidade preditiva. As outras medidas de avaliação mostram também resultados elevados, podendo pois dizer-se que os modelos MIC e MGD-MIC obtêm uma melhoria efectiva na classificação.

Na Tabela 8, quando a separabilidade é baixa, observa-se a seleção do modelo MGD ou MGD-MIC como o melhor modelo, e também neste caso, de classes balanceadas $\phi = Kappa = I.Huberty$, reve-

Tabela 7: Avaliação do desempenho do modelo combinado no caso de classes balanceadas e amostras de pequena dimensão (resultados médios e desvios padrão intra-cenários)

Medidas	$n_1 = n_2 = 30$					
	Separabilidade Baixa			Separabilidade Elevada		
	MIC	MGD	MGD-MIC	MIC	MGD	MGD-MIC
<i>Tx. Bem Class.</i>	0,60 (0,03)	0,61 (0,09)	0,62 (0,07)	0,94 (0,04)	0,89 (0,04)	0,94 (0,04)
<i>Sensibilidade</i>	0,60 (0,07)	0,62 (0,07)	0,66 (0,10)	0,94 (0,04)	0,90 (0,07)	0,96 (0,07)
<i>Especificidade</i>	0,60 (0,08)	0,59 (0,11)	0,59 (0,09)	0,94 (0,05)	0,88 (0,05)	0,92 (0,04)
<i>Precisão</i>	0,60 (0,03)	0,62 (0,09)	0,63 (0,07)	0,94 (0,04)	0,89 (0,04)	0,93 (0,04)
<i>Média Geométrica</i>	0,59 (0,03)	0,60 (0,09)	0,61 (0,08)	0,94 (0,04)	0,89 (0,04)	0,94 (0,04)
<i>Medida F</i>	0,60 (0,04)	0,61 (0,08)	0,64 (0,08)	0,94 (0,04)	0,89 (0,04)	0,94 (0,05)
<i>Tx. Bem Clas. Bal.</i>	0,60 (0,03)	0,61 (0,09)	0,62 (0,07)	0,94 (0,04)	0,89 (0,04)	0,94 (0,04)
<i>Coefficiente ϕ</i>	0,20 (0,05)	0,21 (0,18)	0,25 (0,15)	0,88 (0,08)	0,79 (0,07)	0,89 (0,09)
<i>Estatística Kappa</i>	0,20 (0,05)	0,21 (0,17)	0,25 (0,15)	0,87 (0,08)	0,78 (0,07)	0,88 (0,09)
<i>Índice de Huberty</i>	0,20 (0,05)	0,21 (0,17)	0,25 (0,15)	0,87 (0,08)	0,78 (0,07)	0,88 (0,09)

lando embora uma melhoria efectiva muito baixa. Quando as classes são bem separadas, o modelo MIC obtém resultados muito semelhantes aos de MGD-MIC mas ainda superiores para algumas medidas.

Na Tabela 9, quando se apresentam os resultados para classes não balanceadas e pouco separadas, sobressai, para a maioria das medidas, o modelo combinado. Neste caso, não balanceado, $\phi \neq Kappa \neq$

Tabela 8: Avaliação do desempenho do modelo combinado no caso de classes balanceadas e amostras de grande dimensão (resultados médios e desvios padrão intra-cenários)

Medidas	$n_1 = n_2 = 200$					
	Separabilidade Baixa			Separabilidade Elevada		
	MIC	MGD	MGD-MIC	MIC	MGD	MGD-MIC
<i>Tx. Bem Class.</i>	0,52 (0,02)	0,54 (0,02)	0,54 (0,02)	0,93 (0,02)	0,90 (0,02)	0,92 (0,02)
<i>Sensibilidade</i>	0,51 (0,03)	0,51 (0,05)	0,56 (0,04)	0,91 (0,03)	0,88 (0,05)	0,91 (0,03)
<i>Especificidade</i>	0,54 (0,02)	0,56 (0,04)	0,51 (0,05)	0,94 (0,02)	0,92 (0,02)	0,92 (0,01)
<i>Precisão</i>	0,52 (0,02)	0,53 (0,02)	0,53 (0,02)	0,94 (0,02)	0,91 (0,02)	0,92 (0,01)
<i>Média Geométrica</i>	0,52 (0,02)	0,53 (0,02)	0,53 (0,02)	0,93 (0,02)	0,90 (0,03)	0,92 (0,02)
<i>Medida F</i>	0,52 (0,02)	0,52 (0,03)	0,54 (0,03)	0,92 (0,02)	0,90 (0,03)	0,92 (0,02)
<i>Tx. Bem Clas. Bal.</i>	0,52 (0,02)	0,54 (0,02)	0,54 (0,02)	0,93 (0,02)	0,90 (0,02)	0,92 (0,02)
<i>Coeficiente ϕ</i>	0,05 (0,03)	0,07 (0,04)	0,07 (0,04)	0,85 (0,04)	0,80 (0,05)	0,84 (0,04)
<i>Estatística Kappa</i>	0,05 (0,03)	0,07 (0,04)	0,07 (0,04)	0,85 (0,04)	0,80 (0,05)	0,84 (0,04)
<i>Índice de Huberty</i>	0,05 (0,03)	0,07 (0,04)	0,07 (0,04)	0,85 (0,04)	0,80 (0,05)	0,84 (0,04)

I.Huberty, e os valores obtidos pelo índice de Huberty revelam um pior desempenho do modelo do que se observaria pela aplicação da regra da maioria. Quando as classes são bem separadas, só a *Sensibilidade* não elege o modelo combinado como o melhor modelo.

A análise das classes não balanceadas e amostras de grande dimensão (ver Tabela 10), mostra que, quando pouco separadas, o modelo MGD é eleito por todas as medidas como o melhor modelo, embora

Tabela 9: Avaliação do desempenho do modelo combinado no caso de classes não balanceadas e amostras de pequena dimensão (resultados médios e desvios padrão intra-cenários)

Medidas	$n_1 = 6; n_2 = 54$					
	Separabilidade Baixa			Separabilidade Elevada		
	MIC	MGD	MGD-MIC	MIC	MGD	MGD-MIC
<i>Tx. Bem Class.</i>	0,67 (0,09)	0,62 (0,09)	0,76 (0,07)	0,90 (0,03)	0,80 (0,12)	0,92 (0,02)
<i>Sensibilidade</i>	0,67 (0,19)	0,70 (0,15)	0,63 (0,11)	0,85 (0,17)	0,90 (0,12)	0,85 (0,12)
<i>Especificidade</i>	0,67 (0,10)	0,62 (0,10)	0,78 (0,09)	0,91 (0,04)	0,79 (0,14)	0,93 (0,03)
<i>Precisão</i>	0,20 (0,07)	0,17 (0,05)	0,24 (0,05)	0,54 (0,09)	0,40 (0,13)	0,61 (0,10)
<i>Média Geométrica</i>	0,65 (0,10)	0,59 (0,13)	0,63 (0,12)	0,85 (0,14)	0,83 (0,08)	0,88 (0,06)
<i>Medida F</i>	0,30 (0,09)	0,29 (0,06)	0,35 (0,07)	0,66 (0,08)	0,52 (0,12)	0,69 (0,06)
<i>Tx. Bem Clas. Bal.</i>	0,67 (0,10)	0,66 (0,08)	0,71 (0,06)	0,88 (0,09)	0,85 (0,07)	0,89 (0,05)
<i>Coefficiente ϕ</i>	0,22 (0,13)	0,19 (0,10)	0,28 (0,08)	0,64 (0,09)	0,51 (0,13)	0,67 (0,07)
<i>Estatística Kappa</i>	0,17 (0,10)	0,13 (0,08)	0,24 (0,08)	0,58 (0,12)	0,44 (0,15)	0,65 (0,07)
<i>Índice de Huberty</i>	-2,33 (0,85)	-2,77 (0,88)	-1,37 (0,74)	0,02 (0,34)	-0,98 (1,17)	0,22 (0,19)

com resultados quase sempre iguais aos do modelo combinado. O *Índice de Huberty* volta a revelar um pior desempenho do que se obteria pela regra da maioria. Quando a separabilidade é elevada, o modelo combinado é eleito como o melhor modelo por todas as medidas.

Em qualquer das tabelas de resultados pode notar-se que, o desvio padrão relativo a todas as medidas apresentadas, é sempre extrema-

Tabela 10: Avaliação do desempenho do modelo combinado no caso de classes não balanceadas e amostras de grande dimensão (resultados médios e desvios padrão intra-cenários)

Medidas	$n_1 = 40; n_2 = 360$					
	Separabilidade Baixa			Separabilidade Elevada		
	MIC	MGD	MGD-MIC	MIC	MGD	MGD-MIC
<i>Tx. Bem Class.</i>	0,54 (0,04)	0,56 (0,03)	0,56 (0,04)	0,90 (0,03)	0,89 (0,05)	0,91 (0,03)
<i>Sensibilidade</i>	0,54 (0,06)	0,58 (0,07)	0,57 (0,07)	0,90 (0,06)	0,89 (0,04)	0,91 (0,04)
<i>Especificidade</i>	0,54 (0,05)	0,55 (0,03)	0,55 (0,05)	0,90 (0,03)	0,89 (0,05)	0,91 (0,03)
<i>Precisão</i>	0,12 (0,02)	0,13 (0,02)	0,13 (0,02)	0,52 (0,08)	0,52 (0,12)	0,54 (0,10)
<i>Média Geométrica</i>	0,53 (0,04)	0,56 (0,05)	0,56 (0,04)	0,90 (0,04)	0,89 (0,03)	0,91 (0,03)
<i>Medida F</i>	0,19 (0,03)	0,21 (0,11)	0,21 (0,03)	0,66 (0,08)	0,64 (0,10)	0,67 (0,08)
<i>Tx. Bem Clas. Bal.</i>	0,54 (0,04)	0,57 (0,06)	0,56 (0,04)	0,90 (0,04)	0,89 (0,03)	0,91 (0,03)
<i>Coeficiente ϕ</i>	0,05 (0,05)	0,08 (0,07)	0,08 (0,05)	0,64 (0,08)	0,62 (0,12)	0,65 (0,08)
<i>Estatística Kappa</i>	0,03 (0,03)	0,05 (0,08)	0,05 (0,03)	0,61 (0,09)	0,59 (0,12)	0,62 (0,08)
<i>Índice de Huberty</i>	-3,59 (0,45)	-3,44 (0,72)	-3,44 (0,45)	0,04 (0,32)	-0,09 (0,46)	0,08 (0,31)

mente baixo, próximo de zero, exceto no caso do *Índice de Huberty* em classes não balanceadas.

5 Conclusões

Pensando no objetivo de avaliar o desempenho do modelo combinado comparativamente aos modelos originais, a medida de avaliação

usada não parece influenciar a decisão, revelando o modelo combinado particular interesse em situações com nível de complexidade elevado, nomeadamente com amostras de pequena dimensão. No caso balanceado, o modelo eleito como o melhor é o mesmo quer com a medida tradicional quer com outra medida como a *Taxa de Bem Classificados Balanceada*. No caso não balanceado, com grande desequilíbrio entre a dimensão das classes, o modelo combinado mostra também o seu interesse, mesmo quando a separabilidade é elevada. O modelo MIC revela um bom desempenho quando o nível de complexidade não é demasiado elevado e o modelo MGD só consegue revelar-se superior aos outros dois modelos quando as amostras não têm pequena dimensão e o nível de complexidade não é demasiado elevado. Como esperado, relativamente à comparação entre as medidas de avaliação, as medidas mais usuais e as combinadas mostram resultados muito semelhantes quando as classes têm dimensões pouco desequilibradas. Quando se regista um forte desequilíbrio entre a dimensão das classes, as medidas combinadas revelam, então, o seu interesse. Por outro lado, o *Coefficiente ϕ* , a *Estatística Kappa* e o *Índice de Huberty* fornecem claramente uma informação de carácter diferente sobre o classificador, cuja interpretação precisa de ser mais explorada, provavelmente em aplicações com dados reais. As medidas *Sensibilidade* e *Especificidade* só revelam particular interesse quando, num certo campo de aplicação como, por exemplo, em Medicina, um dos erros de classificação é considerado particularmente importante.

A avaliação dos resultados em Classificação Supervisionada continuará a ser explorada recorrendo quer a dados simulados (considerando um maior número de réplicas em cada cenário e um maior número de cenários) quer a dados reais, particularmente no caso de classes não balanceadas, procurando compreender melhor o interesse de outras medidas de avaliação menos tradicionais, como por exemplo, o *Índice de Huberty*.

Referências

- [1] Bacelar-Nicolau, H. (1985). The affinity coefficient in cluster analysis. *Methods of Operations Research*, 53, 507–512.
- [2] Brito, I., Celeux, G., Sousa Ferreira, A. (2006). Combining methods in supervised classification: A comparative study on discrete and continuous problems. *REVSTAT - Statistical Journal*, 4, 201–225.
- [3] Celeux, G., Nakache, J. P. (1994). *Analyse Discriminante sur Variables Qualitatives*. Celeux, G., Nakache, J.P. (eds.), Polytechnica.
- [4] Ferreira, A. S., Cardoso, M. G. (2013). Evaluating Discriminant Analysis Results. In: Lita da Silva J., Caeiro F., Natário I., Braumann C. (eds.): *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and and Other Statistical Applications. Studies in Theoretical and Applied Statistics*, 155–162, Springer, Berlin, Heidelberg.
- [5] Goldstein, M., Dillon, W.R. (1978). *Discrete Discriminant Analysis*. Wiley and Sons.
- [6] Ho, T.K., Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 289–300.
- [7] Marques, A. (2014). *Análise Discriminante sobre Variáveis Qualitativas*. Tese de Doutoramento, ISCTE - Instituto Universitário de Lisboa.
- [8] Marques, A., Sousa Ferreira, A., Cardoso, M. (2017). Performance of Combined Models in Discrete Binary Classification. *Methodology* 13(1), 23–37.
- [9] Paik, H. (1998). The effect of prior probability on skill in two-group discriminant analysis. *Quality and Quantity*, 32(2), 201–211.
- [10] Santafe, G., Inza, I., Lozano, J.A. (2015). Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44(4), 467–508.
- [11] Sousa Ferreira, A. (2000). *Combinação de Modelos em Análise Discriminante sobre Variáveis Qualitativas*. Tese de Doutoramento, Universidade Nova de Lisboa.

Comparação bayesiana de testes de diagnóstico com dados densamente omissos ao acaso

Carlos Daniel Paulino

Centro de Estatística e Aplicações & IST, Universidade de Lisboa,
dpaulino@math.ist.utl.pt

Giovani L. Silva

Dep. Matemática, Instituto Superior Técnico & CEAUL, Universidade de Lisboa, *giovani.silva@tecnico.ulisboa.pt*

Palavras-chave: Dados categorizados omissos ao acaso; Medidas de acurácia; Metodologia bayesiana; Método MCMC; Distribuição Dirichlet generalizada.

Resumo: Este trabalho é uma sequência de um artigo (Poletto *et al.* [7]) sobre comparação de testes, assente num conhecido padrão de ouro, através das usuais medidas de acurácia, por meio de métodos frequentistas num quadro de substancial omissão de dados segundo um processo não informativo. Este artigo passa a adotar uma abordagem bayesiana por se entender mais adequada para lidar com a escassez da subamostra completa e a incompletude do grosso dos dados, sem recorrer a argumentos válidos para grandes amostras. Computacionalmente propõe-se que a análise recorra a um método de Monte Carlo com ampliação de dados, reduzindo tanto quanto possível a fase de imputação. Em cada passo *a posteriori*, após fácil simulação de apropriadas variáveis latentes, o parâmetro de interesse é simulado diretamente à custa de distribuições Dirichlet.

1 Introdução

Poleto *et al.* [7] procedem a uma comparação prática de testes de diagnóstico binário na presença de dados faltantes através de abordagens frequencistas apoiadas em resultados para grandes amostras. Nelas se incluem análises simplistas de subconjuntos mais ou menos restritivos dos dados observados e que se mostra serem inferiores à análise integral de todos os dados baseada num processo gerador da omissão consistente com as causas desta. Esta última análise radicada num mecanismo de omissão não informativa pode ser executada no software ACD elaborada pelos autores supracitados e disponibilizada no repositório CRAN (*vide* Poleto *et al.* [8]). A sua fundamentação teórica é descrita designadamente em Poleto *et al.* [9].

Neste trabalho usa-se um conjunto de dados do mesmo estudo analisado no artigo acima referido que envolveu $N=219$ pacientes, submetidas para deteção de endometriose (retrocervical) ao procedimento (D) de laparoscopia (considerado como padrão de ouro) e com resultado devidamente registado. Os testes de natureza não invasiva a comparar aqui respeitam apenas a ressonância magnética retrocervical (MR) e a ecocolonoscopia (EC) e os seus resultados não foram observados para um número significativo de pacientes – apenas se conheceu o resultado de ambos os testes para cerca de 6% das pacientes. A ocorrência de omissão deveu-se à indisponibilidade dos equipamentos no momento da comparência das unidades amostrais. Os dados observados são reproduzidos na Tabela 1.

Dada a substancial incompletude classificativa na amostra selecionada e a exiguidade da subamostra completamente observada, o objetivo aqui é proceder a uma abordagem bayesiana da totalidade do que foi observado, propondo uma estratégia computacional eficiente para a realização das inferências de interesse. Estas dizem respeito às usuais medidas de acurácia dos testes em comparação conhecidas como sensibilidade, especificidade e valores preditivos positivo e negativo.

Tabela 1: Frequências observadas de pacientes

Ressonância magnética (<i>MR</i>)	Ecocolonos- copia (<i>EC</i>)	Endometriose (<i>D</i>)	
		—	+
—	—	6	1
	+	1	2
	omisso	51	22
+	—	0	1
	+	0	2
	omisso	5	13
omisso	—	3	5
	+	3	6
	omisso	53	45

2 Modelação estatística

O facto de a ocorrência de falhas em unidades amostrais não se dever a qualquer delineamento prefixado conduz a que se deva introduzir uma variável qualitativa adicional, diga-se W , que indique os distintos padrões de omissão que se verificaram. Por exemplo, tomando $W=1$ para a não omissão, $W=2(3)$ para a omissão do resultado apenas do teste $EC(MR)$ e $W=4$ para a omissão do resultado de ambos os testes.

A tabela ampliada $2^3 \times 4$ para (MR, EC, D, W) está naturalmente recheada de frequências desconhecidas. Denotando por M o vetor das frequências $m_{ijk r}$, $r = 1, 2, 3, 4$; $i, j, k = 1, 2$ com 1(2) indicando resultado positivo (negativo), admite-se para ele uma distribuição Multinomial com parâmetro probabilístico $\gamma = (\gamma_{ijk r})$ que se pode fatorizar do seguinte modo:

$$\begin{aligned} \gamma_{ijk r} &= P(MR=i, EC=j, D=k) P(W=r | MR=i, EC=j, D=k) \\ &\equiv \theta_{ijk} \times \lambda_{r(ijk)}. \end{aligned}$$

O parâmetro $\theta = (\theta_{ijk})$, com $\sum_{i,j,k} \theta_{ijk} = 1$, caracteriza o processo

marginal de classificação segundo (MR,EC,D) e $\lambda = (\lambda_{r(ijk)})$, com $\sum_r \lambda_{r(ijk)} = 1, \forall i, j, k$, o processo condicional de omissão. Note-se que a situação trivial de ausência sistemática de omissão resulta de se fazer $\lambda_{r(ijk)} = 0, r \neq 1$, originando para $M = (m_{ijk1})$ o modelo padrão Multinomial, $M_7(N, \theta)$, em que o índice 7 neste seu símbolo indica a dimensionalidade do respetivo vetor aleatório.

Os dados realmente observados são facilmente expressos em termos de componentes de M por $\mathcal{D}_0 = \{n_{ijk}, s_{ik}, q_{jk}, p_k\}$, com $n_{ijk} = m_{ijk1}$, $s_{ik} = m_{i\bullet k2}$, $q_{jk} = m_{Cjk3}$ e $p_k = m_{\bullet\bullet k4}$ (*vide nota*¹). A previsível sobreparametrização do modelo para os dados completados M costuma ser erradicada por restrições sobre λ decorrentes de informação (ou suposição) sobre causas possíveis da omissão.

Impondo a condição de $\{\lambda_{r(ijk)}\}$ não dependerem do que não foi observado, obtém-se o chamado processo de omissão ao acaso (MAR) definido por

$$\lambda_{4(ijk)} = \delta_k, \quad \lambda_{3(ijk)} = \beta_{jk}, \quad \lambda_{2(ijk)} = \alpha_{ik}, \quad \lambda_{1(ijk)} = \eta_{ijk}, \forall i, j, k,$$

o qual conduz a um modelo saturado para o vetor de frequências observadas com a verosimilhança Multinomial fatorizável do seguinte modo

$$\begin{aligned} L(\theta, \lambda^* | \mathcal{D}_0) &\propto \left[\prod_{i,j,k} \theta_{ijk}^{n_{ijk}} \prod_{i,k} \theta_{i\bullet k}^{s_{ik}} \prod_{j,k} \theta_{\bullet j k}^{q_{jk}} \prod_k \theta_{\bullet\bullet k}^{p_k} \right] \times \\ &\times \left[\prod_{i,j,k} \eta_{ijk}^{m_{ijk1}} \prod_{i,k} \alpha_{ik}^{s_{ik}} \prod_{j,k} \beta_{jk}^{q_{jk}} \prod_k \delta_k^{p_k} \right] \equiv L(\theta | \mathcal{D}_0) \times L(\lambda^* | \mathcal{D}_0), \end{aligned}$$

em que λ^* denota os elementos distintos de λ sob MAR.

Um caso especial deste processo é obtido quando $\{\lambda_{r(ijk)}\}$ não dependem também do que foi observado, ou seja $\forall i, j, k$,

$$\lambda_{4(ijk)} = \delta, \quad \lambda_{3(ijk)} = \beta, \quad \lambda_{2(ijk)} = \alpha, \quad \lambda_{1(ijk)} = \eta = 1 - (\alpha + \beta + \delta),$$

¹Em conformidade com uma notação usual, as quantidades indexadas com algum símbolo \bullet indicam ser somas para todos os valores do índice substituído por tal símbolo. A título exemplificativo, a tabela das frequências observadas indica que $s_{11} = 13$, $s_{12} = 5$, $s_{21} = 22$ e $s_{22} = 51$.

sendo conhecido como o processo de omissão completamente ao acaso (MCAR). Aqui, sendo $\lambda_* = (\alpha, \beta, \delta)$, tem-se $L(\theta, \lambda_* | \mathcal{D}_0) = L(\theta | \mathcal{D}_0) \times L(\lambda_* | \mathcal{D}_0)$, em que o 2º fator é o núcleo da distribuição Multinomial amostral de $\{N_r \equiv m_{\bullet\bullet\bullet r}\}$. Isto implica que $L(\theta | \mathcal{D}_0)$ passe a ser interpretado como o núcleo de uma distribuição Produto de Multinomiais condicional para as frequências observadas dados os totais dos padrões de omissão N_r . Este mecanismo de omissão é consistente com a informação de que a omissão por não realização de testes foi devida à indisponibilidade do equipamento aquando da comparência de algumas pacientes.

Como os parâmetros de interesse são função de θ , não há diferença entre os dois processos de omissão no que concerne às inferências bayesianas pretendidas, desde que a priori θ seja independente do parâmetro perturbador (probabilidades condicionais de omissão). Note-se que as funções paramétricas de interesse são definidas por

$$\begin{aligned} Sens(MR) &= P(MR=+|D=+) = \frac{\theta_{1\bullet 1}}{\theta_{\bullet\bullet 1}}, \quad Sens(EC) = P(EC=+|D=+) = \frac{\theta_{\bullet 11}}{\theta_{\bullet\bullet 1}}, \\ Spec(MR) &= P(MR=-|D=-) = \frac{\theta_{2\bullet 2}}{\theta_{\bullet\bullet 2}}, \quad Spec(EC) = P(EC=-|D=-) = \frac{\theta_{\bullet 22}}{\theta_{\bullet\bullet 2}}, \\ PPV(MR) &= P(D=+|MR=+) = \frac{\theta_{1\bullet 1}}{\theta_{1\bullet\bullet}}, \quad PPV_{(EC)} = P(D=+|EC=+) = \frac{\theta_{\bullet 11}}{\theta_{\bullet 1\bullet}}, \\ NPV(MR) &= P(D=-|MR=-) = \frac{\theta_{2\bullet 2}}{\theta_{2\bullet\bullet}}, \quad NPV_{(EC)} = P(D=-|EC=-) = \frac{\theta_{\bullet 22}}{\theta_{\bullet 2\bullet}}. \end{aligned}$$

Para os objetivos inferenciais vai considerar-se para distribuição *a priori* de θ um membro da família Dirichlet denotada pelo símbolo $D_7(b)$, $b = (b_{ijk})$ com $b_{ijk} > 0$, que represente de algum modo o grau de vaguidade da informação *a priori* que se pretende considerar na análise. Na Secção 4 usou-se a distribuição Uniforme no simplex heptadimensional.

Dada a expressão de $L(\theta | \mathcal{D}_0)$, a distribuição *a posteriori* de θ apresenta o núcleo

$$h(\theta | \mathcal{D}_0) \propto \prod_{i,j,k} (\theta_{ijk})^{n_{ijk} + b_{ijk} - 1} \prod_{i,k} \theta_{i\bullet k}^{s_{ik}} \prod_{j,k} \theta_{\bullet jk}^{q_{jk}} \prod_k \theta_{\bullet\bullet k}^{p_k}.$$

A forma deste núcleo evidencia que se trata de uma distribuição

Dirichlet generalizada estudada em Dickey[2]. A sua constante normalizadora, definida em termos do valor esperado sob $D_7(n+b)$, $n = (n_{ijk})$, de produtos de várias potências de somas de elementos de θ agrupados em três fatores, não é expressável em forma fechada.

Note-se que esses três fatores são potências de probabilidades das partes de três partições do conjunto $\{(i,j,k), i,j,k = 1,2\}$ de categorias correspondentes à omissão separada de EC (\mathcal{P}_2) e de MR (\mathcal{P}_3) e simultânea dos dois testes (\mathcal{P}_4). Nelas é possível visualizar dois pares de partições sucessivamente encaixadas na partição mais fina, $\mathcal{P}_1 = \{\{(i,j,k)\}, i,j,k = 1,2\}$, que está associada à ausência de omissão.

A forma distribucional do modelo bayesiano para θ dadas as frequências observadas² deixa antever que a aplicação direta de métodos MCMC pelo software bayesiano mais difundido (BUGS) está longe de surtir os desejáveis efeitos.

3 Método computacional

A forma de $L(\theta|\mathcal{D}_0)$ mostra que a redução do número de fatores envolvendo as frequências ligadas às unidades que sofreram omissão pode conduzir a uma verosimilhança analiticamente mais tratável. É o que acontece em qualquer das seguintes situações: ausência de qualquer omissão, omissão ligada à classificação numa única partição de categorias ou numa sequência de duas partições encaixadas.

A obtenção de uma verosimilhança com tal estrutura no atual quadro de registo de quatro padrões de omissão consegue-se ampliando \mathcal{D}_0 a um apropriado vetor z de frequências hipotéticas m_{ijk_r} para $r \neq 1$. Optando por reduzir o número dessas variáveis não observadas, tome-se por exemplo $z = (z_{ijk})$ relativo à discriminação das frequências das unidades com omissão em MR, *i.e.* $z_{ijk} = m_{ijk3}$. Pelas propriedades da distribuição Multinomial de M , tem-se sob

²No caso de se pretender inferir também sobre as probabilidades condicionais de omissão, o uso no quadro MCAR de uma distribuição *a priori* Dirichlet para λ_* conduz a uma distribuição *a posteriori* da mesma família para ele.

MAR que

$$L(\theta, \lambda^* | \mathcal{D}_0, z) = L(\theta | \mathcal{D}_0, z) \times L(\lambda^* | \mathcal{D}_0),$$

em que

$$L(\theta | \mathcal{D}_0, z) \propto \prod_{i,j,k} (\theta_{ijk})^{n_{ijk} + z_{ijk}} \prod_{i,k} \theta_{i\bullet k}^{s_{ik}} \prod_k \theta_{\bullet\bullet k}^{p_k}$$

já tem uma estrutura envolvendo os padrões de omissão nas partições sucessivamente encaixadas \mathcal{P}_4 , \mathcal{P}_2 e \mathcal{P}_1 .

Note-se ainda que $L(\theta | \mathcal{D}_0, z) = L(\theta | \mathcal{D}_0) f(z|q, \theta)$, em que o 2º fator traduz o produto das distribuições

$$z_{1jk} | q_{jk}, \theta \underset{ind}{\sim} M_1(q_{jk}, \theta_{1jk} / \theta_{\bullet jk}), j, k = 1, 2,$$

como consequência de se verificar $z | \theta, \lambda \sim M_8(N, \{\theta_{ijk} \beta_{jk}\})$ e $q = (q_{jk} \equiv z_{\bullet jk}) | \theta, \lambda \sim M_4(N, \{\theta_{\bullet jk} \beta_{jk}\})$.

A distribuição *a posteriori* conjunta das quantidades não observadas θ e z é então dada por

$$h(\theta, z | \mathcal{D}_0) \propto h(\theta) L(\theta | \mathcal{D}_0, z) = h(\theta | \mathcal{D}_0) f(z | q, \theta),$$

traduzindo a desmarginalização da distribuição *a posteriori* de interesse $h(\theta | \mathcal{D}_0)$.

A distribuição *a posteriori* de θ condicional aos dados ampliados (\mathcal{D}_0, z) ,

$$h(\theta | \mathcal{D}_0, z) \propto \prod_{i,j,k} (\theta_{ijk})^{b_{ijk} + n_{ijk} + z_{ijk} - 1} \prod_{i,k} \theta_{i\bullet k}^{s_{ik}} \prod_k \theta_{\bullet\bullet k}^{p_k}$$

é uma outra distribuição Dirichlet generalizada mas com os fatores relativos às frequências observadas $s = (s_{ik})$ e $p = (p_k)$ agrupados segundo a sequência das partições encaixadas \mathcal{P}_4 e \mathcal{P}_2 , o que implica uma expressão explícita da constante normalizadora em termos de funções Beta completas (Dickey *et al.* [3], Jiang *et al.* [4]). A

estrutura desta distribuição permite que ela apresente uma caracterização simpática em termos de distribuições Dirichlet independentes, tomando em consideração a seguinte reparametrização de θ :

$$\theta_{ijk} = \frac{\theta_{ijk}}{\theta_{i\bullet k}} \times \frac{\theta_{i\bullet k}}{\theta_{\bullet\bullet k}} \times \theta_{\bullet\bullet k} \equiv \rho_j^{ik} \times \varepsilon_i^k \times \phi_k.$$

A referida caracterização de $h(\theta|\mathcal{D}_0, z)$ é tal que condicionalmente a (\mathcal{D}_0, z)

$$\{\theta_{ijk}\} : \begin{cases} \rho^{ik} = (\rho_j^{ik}, j = 1, 2) \sim D_1(b_{ijk} + n_{ijk} + z_{ijk}, j = 1, 2) \\ \varepsilon^k = (\varepsilon_i^k, i = 1, 2) \sim D_1(b_{i\bullet k} + n_{i\bullet k} + z_{i\bullet k} + s_{ik}, i = 1, 2) \\ \phi = (\phi_k, k = 1, 2) \sim D_1(b_{\bullet\bullet k} + n_{\bullet\bullet k} + z_{\bullet\bullet k} + s_{\bullet k} + p_k, k = 1, 2). \end{cases}$$

A fácil simulação pelas rotinas disponíveis das distribuições condicionais de $\theta|\mathcal{D}_0, z$ e de $z|q, \theta$ em ciclos de dois passos permite obter uma amostra simulada da distribuição-alvo $h(\theta|\mathcal{D}_0)$, após convergência do denominado algoritmo de ampliação de dados em cadeia (Tanner e Wong [10]), com base na qual se determinam as inferências de interesse. Este algoritmo é descrito como segue:

- **Passo de imputação:** Partindo de $\theta^{(t)}$ simula-se $\{z_{ijk}^{(t+1)}\}$ de

$$z_{1jk}|q_{jk}, \theta^{(t)} \sim M_1(q_{jk}, \frac{\theta_{1jk}}{\theta_{\bullet jk}^{(t+1)}}), z_{2jk} = q_{jk} - z_{1jk}, j, k = 1, 2;$$

- **Passo *a posteriori*:** Calcula-se $\theta^{(t+1)}$ das equações $\theta_{ijk} = \rho_j^{ik} \varepsilon_i^k \phi_k$ com os seus fatores simulados das distribuições Beta acima usando $z_{ijk}^{(t+1)}, \forall i, j, k$.

Repetindo o esquema cíclico, a amostra retida de θ , bem como de qualquer função $\psi(\theta)$, após convergência respeita às suas distribuições *a posteriori* dado \mathcal{D}_0 . A natureza das fontes simuladoras permite visualizar este amostrador como um algoritmo Gibbs para

amostragem da distribuição ampliada $h(\theta, z | \mathcal{D}_0)$. Para mais detalhes, veja-se *e.g.* Amaral Turkman e Paulino [1] ou Paulino *et al.* [6].

Tendo em conta que a distribuição *a posteriori* para θ do tipo da de $h(\theta | \mathcal{D}_0, z)$ apresenta momentos (e outros resumos pontuais) em forma fechada, um bom valor para iniciar o algoritmo é tomar por exemplo $\theta^{(0)}$ como a média *a posteriori* da distribuição Dirichlet generalizada para θ dado $\mathcal{D}_0 - \{q_{jk}\}$, denotada por $\bar{\theta}_{ijk}$ e dada explicitamente por

$$\left(b_{ijk} + n_{ijk} + s_{ik} \frac{b_{ijk} + n_{ijk}}{b_{i\bullet k} + n_{i\bullet k}} + p_k \frac{b_{ijk} + n_{ijk}}{b_{i\bullet k} + n_{i\bullet k}} \frac{b_{i\bullet k} + n_{i\bullet k} + s_{ik}}{b_{\bullet\bullet k} + n_{\bullet\bullet k} + s_{\bullet k}} \right) (b_{\bullet} + N_-)^{-1}$$

com $b_{\bullet} = \sum_{i,j,k} b_{ijk}$ e $N_- = N - q_{\bullet\bullet} = 202$.

4 Análise de resultados

A implementação computacional do algoritmo atrás descrito fez-se através da criação de um programa *ad-hoc* no software R. A convergência das cadeias dos elementos de θ atingiu-se rapidamente como se revela pelos resultados dos testes de convergência e dos gráficos, omitidos por motivos de espaço, de traços e das autocorrelações (estas desaparecem num ápice).

O gráfico de evolução das médias empíricas mostra que a estabilidade é atingida ao fim de poucas centenas de iterações, independentemente do valor inicial – veja-se para exemplificação a Figura 1 exibindo os gráficos correspondentes a θ_{112} , θ_{211} e θ_{222} e a duas medidas de acurácia de testes, quando a cadeia se iniciou com equiprobabilidade no seio de θ .³ A amostra simulada para base das inferências reuniu 4000 iteradas escolhidas sem qualquer desbaste após um período de aquecimento de tamanho 1000.

³O comportamento eficiente desta cadeia está em completa dissonância quando se compara com o obtido por Metropolis-Hastings (JAGS) diretamente do modelo $h(\theta | \mathcal{D}_0)$.

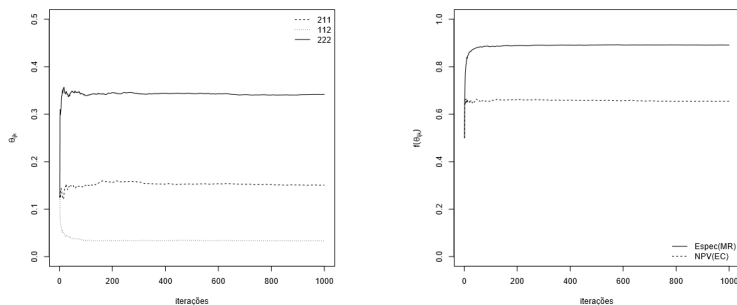


Figura 1: Gráficos das médias ergódicas para $\{\theta_{ijk}\}$ (esquerda) e para especificidade (MR) e NPV (EC) (direita).

A Tabela 2 exibe as médias *a posteriori* dos elementos de θ , bem como os valores iniciais propostos, calculados como se sugeriu anteriormente, cuja proximidade com as estimativas finais era expectável.

Tabela 2: Médias *a posteriori* (valores iniciais) de $\{\theta_{ijk}\}$

MR	EC	D	
		$k = 1$	$k = 2$
$i = 1$	$j = 1$	0.105 (0.103)	0.033 (0.030)
	$j = 2$	0.074 (0.069)	0.026 (0.030)
$i = 2$	$j = 1$	0.150 (0.154)	0.153 (0.114)
	$j = 2$	0.116 (0.103)	0.342 (0.398)

A Figura 2 apresenta as densidades das diferenças entre os testes MR e EC da especificidade e da valor preditivo negativo. Elas evidenciam, por um lado, uma superioridade razoável de MR sobre EC em termos da especificidade e, por outro, a equivalência dos dois testes no que respeita ao NPV.

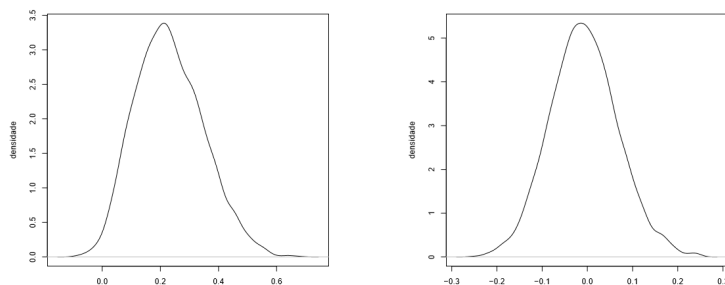


Figura 2: Densidades *a posteriori* de DIF(Espec) e DIF(NPV).

Estas conclusões também são patentes na Tabela 3 ao olhar para as estimativas pontuais e intervalares das diferenças das medidas de acurácia entre os dois testes, juntamente com os respetivos níveis de plausibilidade relativa *a posteriori*⁴ de uma diferença nula. Com base nestes dados, no global não parece que os dois testes e, em especial o EC, sejam concorrentes à altura do teste perfeito.

Tabela 3: Médias *a posteriori* (IC HPD 95%) de funções de interesse

	Sens	Spec	PPV	NPV
MR	0.40	0.89	0.75	0.65
EC	0.57	0.66	0.58	0.66
DIF	-0.17	0.23	0.17	-0.008
	(-0.42,0.06)	(0.02,0.46)	(-0.07,0.39)	(-0.15,0.15)
	0.1958*	0.0372*	0.1566*	0.8402*

(*) Nível de plausibilidade relativa *a posteriori* para DIF=0

⁴Para definição deste conceito veja-se *e.g.* Paulino *et al.* [6], Sec. 3.4.2.

5 Conclusões

Este artigo sobre testes de diagnóstico debruça-se sobre um problema específico de dados categorizados referentes ao cruzamento de três variáveis respostas binárias, onde a larga maioria das unidades amostrais apresenta uma classificação incompleta. Devido aos diversos padrões registados de omissão não informativa, a distribuição *a posteriori* do vetor de probabilidades de categorização não é expressável inteiramente em forma fechada nem facilmente simulável.

O método computacional proposto configura um esquema de Monte Carlo iterativo com ampliação de dados, onde as frequências latentes são convenientemente escolhidas de modo que as decorrentes distribuições *a posteriori* para os dados ampliados propiciam uma simulação direta. As vantagens deste método residem na sua assinalável eficiência, demonstrada na aplicação ao caso em estudo, e na sua adaptabilidade a problemas análogos de tabelas de contingência multidimensionais com padrões variáveis de incompletude ao acaso.

Agradecimentos

Este trabalho foi parcialmente financiado por intermédio do CEAUL pela FCT através do projeto UID/MAT/UI0006/2013. Estamos gratos ao colega Paulo Soares pela cedência da sua rotina R para cálculo dos níveis de plausibilidade relativa *a posteriori* de hipóteses paramétricas. Agradecemos ainda aos dois avaliadores do trabalho submetido pelas sugestões formuladas que redundaram numa versão melhorada do artigo.

Referências

- [1] Amaral Turkman, M.A. e Paulino, C.D. (2015). *Estatística Bayesiana Computacional – uma introdução*. Edições SPE, Lisboa.

- [2] Dickey, J.M. (1983). Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *J. Amer. Statist. Assoc.*, 78, 628-637.
- [3] Dickey, J.M., Jiang, J.M. e Kadane, J.B. (1987). Bayesian methods for censored categorical data. *J. Amer. Statist. Assoc.*, 82, 773-781.
- [4] Jiang, J.M., Kadane, J.B. e Dickey, J.M. (1992). Computation of Carlson's multiple hypergeometric functions \mathcal{R} for Bayesian applications. *J. Statist. and Comput. Graphics*, 1, 231-251.
- [5] Paulino, C.D. (1988). *Análise de Dados Categorizados Incompletos: Fundamentos, Métodos e Aplicações*. Tese de doutoramento, IME-Universidade de São Paulo.
- [6] Paulino, C.D., Amaral Turkman, M.A., Murteira, B. e Silva, G.L. (2018). *Estatística Bayesiana*, 2ªed.. Fundação Calouste Gulbenkian, Lisboa.
- [7] Poleto, F.Z., Singer, J.M. e Paulino, C.D. (2011). Comparing diagnostic tests with missing data. *J. Applied Statistics*, 38, 1207-1222.
- [8] Poleto, F., Singer, J., Paulino, C.D., Correa, F. e Jelihovschi, E. (2013). *Package ACD: Categorical data analysis with complete or missing responses*. Version 1.5. CRAN (Comprehensive R Archive Network).
- [9] Poleto, F.Z., Singer, J.M. e Paulino, C.D. (2014). A product-multinomial framework for categorical data analysis with missing responses. *Brazilian Journal of Probability and Statistics* 28, 1, 109-139.
- [10] Tanner, M.A. e Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.*, 82, 528-550.

O critério *Minimum Message Length* na estimação de modelos de mistura sobre dados mistos

Cláudia Silvestre

Escola Superior de Comunicação Social-Instituto Politécnico de Lisboa, csilvestre@escs.ipl.pt

Margarida G. M. S. Cardoso

Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, margarida.cardoso@iscte.pt

Mário A. T. Figueiredo

Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Portugal, mario.figueiredo@lx.it.pt

Palavras-chave: Classificação não supervisionada; Análise de Agrupamento; Modelos de Mistura Finita; Dados Mistos; Minimum Message Length.

Resumo: Neste trabalho propomos uma nova variante do algoritmo *Expectation-Maximization* para agrupar dados mistos que simultaneamente estima o número de grupos. Recorremos aos modelos de mistura finita, pressupondo que os dados categoriais são modelados por distribuições multinomiais e os métricos por distribuições gaussianas. Para estimar o número de componentes de mistura baseamos-nos no critério *Minimum Message Length*. O desempenho do algoritmo proposto, designado por EM-MML-mix, é comparado com o de outros critérios usados frequentemente para a seleção de modelos de mistura. Desta análise comparativa, realizada sobre dados simulados e sobre um conjunto de dados reais provenientes do *European Social Survey*, salienta-se o reduzido tempo de computação para a obtenção da solução mediante a metodologia proposta.

1 Introdução

O agrupamento sobre dados mistos é um problema prático comum, nomeadamente no âmbito das ciências sociais. Este pode referir-se, por exemplo, à constituição de segmentos homogêneos de indivíduos, considerando as suas características métricas ou qualitativas. As abordagens metodológicas a este problema têm sido diversas. Por exemplo, Chiu et al. [8] propõem um algoritmo incremental e Ahmad e Dey [1] propõem um novo algoritmo K-Médias, ambos capazes de lidar com dados métricos e categoriais.

No âmbito do agrupamento com modelos de mistura finita, uma primeira proposta considerando dados mistos deve-se a Everitt [10]. A vantagem desta abordagem para segmentação, reside na sua capacidade de analisar diversos tipos de variáveis, de modelar relações entre elas, de integrar diversos critérios de seleção dos modelos e ainda de selecionar o número de segmentos (componentes da mistura).

Um modelo de mistura finita considera uma distribuição conjunta para as variáveis base de segmentação como uma soma ponderada de distribuições intra-segmentos, atendendo à natureza diversa dos atributos. A sua estimação viabiliza a construção de uma estrutura probabilística de segmentos e, em simultâneo, a obtenção de estimativas dos parâmetros distribucionais intra-segmentos. Neste âmbito, Hunt e Jorgensen [13] modelam a distribuição conjunta de uma variável categorial e de multinormais, permitindo, nestas últimas, que as médias dependam das categorias da variável qualitativa (sendo as covariâncias comuns). Outros trabalhos integram, nos modelos de mistura finita, a modelação conjunta de variáveis mistas considerando diversas distribuições, admitindo correlações intra-grupos de variáveis métricas ou mesmo de variáveis métricas contínuas (por exemplo, [20] e [15]). O critério que habitualmente orienta a estimação destes modelos é o da máxima verosimilhança. No entanto, incorporando informação *a priori*, podem também adotar-se métodos bayesianos.

Neste trabalho, consideramos o agrupamento de dados mistos, usando um modelo de mistura e propondo o uso do critério *Minimum Mes-*

sage Length (MML) [21] para a sua estimação. Este critério advém da teoria da informação, considerando como modelo mais adequado aquele que permite uma descrição mais sucinta das observações. Figueiredo e Jain [11] foram pioneiros na utilização deste critério para estimação de misturas de gaussianas e uma primeira proposta para a utilização do MML em misturas de multinomiais foi proposta por Silvestre et al. [19]. Este critério também foi usado em agrupamento de dados *fuzzy* em [16], onde os autores consideraram misturas de gaussianas e usaram o MML para estimar as variáveis relevantes e identificar o número de componentes de mistura.

A presente análise integra dados mistos considerando uma mistura de gaussianas e multinomiais, bem como um algoritmo que é uma variante do conhecido *Expectation-Maximization* (EM). A metodologia é testada comparativamente com critérios comuns para a seleção de modelos de mistura, nomeadamente o Integrated Completed Likelihood, o qual é particularmente adequado neste contexto [12]. A análise é efetuada sobre dados sintéticos e um conjunto de dados reais (provenientes do *European Social Survey*). São feitas análises comparativas quanto ao tempo de computação, à qualidade do agrupamento obtido e à robustez, relativamente a diferentes processos de inicialização.

2 Metodologia

Em muitos dos trabalhos propostos a escolha do número de grupos é feita *a posteriori*. Por exemplo, nos métodos hierárquicos, a escolha do número de grupos é feita após o agrupamento, recorrendo aos correspondentes dendrogramas. Os critérios baseados na verossimilhança, habitualmente combinados com a estimação de modelos de mistura finita, também necessitam que o agrupamento seja feito previamente. Entre estes critérios, são comuns os seguintes: *Bayesian Information Criterion* (BIC) [17], *Akaike Information Criterion* (AIC) [2] e suas variantes [5, 6] e *Integrated Complete Likelihood* (ICL)[4]. No uso destes critérios, o agrupamento é feito para dife-

rentes números de grupos e escolhe-se a solução que corresponde ao melhor valor do critério usado. A metodologia que se propõe incorpora a determinação do número de grupos na estimação do modelo de mistura.

2.1 Modelos de mistura finita

Os modelos de mistura finita têm uma longa tradição em agrupamento; e.g., Wedel e Kamakura [22] referem o seu uso no âmbito de aplicações em marketing. A sua natureza probabilística/estatística tem várias vantagens importantes. Nomeadamente, a possibilidade de se modelar dados de diferentes naturezas e de se abordar formalmente a estimação do número de grupos.

Seja $\mathbf{Y} = \{\underline{y}_i, i = 1, \dots, n\}$ uma amostra aleatória de n observações independentes de $\underline{Y} = [Y_1, \dots, Y_D]'$. A ideia base dos modelos de mistura finita é considerar a distribuição conjunta para as variáveis base de segmentação \underline{Y} como sendo uma soma ponderada de distribuições intra-segmentos,

$$f(\underline{y}|\Theta) = \sum_{k=1}^K \alpha_k f(\underline{y}|\underline{\theta}_k),$$

onde $\Theta = \{\underline{\theta}_1, \dots, \underline{\theta}_K, \alpha_1, \dots, \alpha_K\}$ é o conjunto de todos os parâmetros do modelo, K o número total de segmentos e $\underline{\theta}_k$ representa o conjunto dos parâmetros distribucionais do k -ésimo segmento (componente de mistura). Os pesos $\alpha_1, \dots, \alpha_K$ são as probabilidades de cada segmento, pelo que $\alpha_k \geq 0$, para $k = 1, \dots, K$ e $\sum_{k=1}^K \alpha_k = 1$. Em agrupamento, a componente de mistura de onde provém cada uma das observações é desconhecida, por isso, pode dizer-se que os dados observados, \mathbf{Y} , são dados incompletos. Essa informação em falta é usualmente designada por \mathbf{Z} : $\mathbf{Z} = \{\underline{z}_1, \dots, \underline{z}_n\}$ onde $\underline{z}_i = [z_{i1}, \dots, z_{iK}]'$ e z_{ik} é um indicador binário que toma o valor 1 se a observação \underline{y}_i foi gerada pela k -ésima componente e 0 caso contrário. É habitual assumir-se que $\{\underline{z}_i, i = 1, \dots, n\}$ são i.i.d. e

que seguem uma distribuição multinomial com K categorias e probabilidades $\{\alpha_1, \dots, \alpha_K\}$. Assim, o logaritmo da verosimilhança dos dados completos, (\mathbf{Y}, \mathbf{Z}) , é dado por

$$\log f(\mathbf{Y}, \mathbf{Z} | \Theta) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \log \left[\alpha_k f(\underline{y}_i | \underline{\theta}_k) \right].$$

Neste trabalho pretendemos agrupar/segmentar dados mistos, ou seja, de natureza categorial e métrica. Consideremos que \underline{Y} tem M variáveis categoriais que serão modeladas por distribuições multinomiais e G variáveis métricas que serão modeladas por distribuições gaussianas, tal que $M + G = D$. Assumindo que as variáveis são condicionalmente independentes, o logaritmo da verosimilhança para os dados completos é dado por:

$$\log f(\mathbf{Y}, \mathbf{Z} | \Theta) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \log \left[\alpha_k \prod_{m=1}^M f(\underline{y}_{im} | \underline{\theta}_{km}) \prod_{g=1}^G f(\underline{y}_{ig} | \underline{\theta}_{kg}) \right].$$

Para se obter as estimativas de máxima verosimilhança é habitual recorrer-se ao algoritmo *Expectation Maximization* (EM) [9].

2.2 O algoritmo EM

O algoritmo EM é um algoritmo iterativo que é frequentemente usado quando se pretende obter as estimativas de máxima verosimilhança (ML) ou o máximo *a posteriori* (MAP) na presença de dados incompletos. Um problema bem conhecido deste algoritmo é conduzir a um máximo local (não necessariamente ao global) da função de verosimilhança. Uma forma habitual de ultrapassar este problema consiste em calcular várias estimativas obtidas com condições iniciais diferentes, escolhendo-se para solução final aquela que apresentar valor mais elevado da função de verosimilhança.

O algoritmo EM alterna entre dois passos:

Passo E: Calcula o valor esperado do logaritmo da verosimilhança completa, condicional aos dados observados

$$E \left[\log f(\mathbf{Y}, \mathbf{Z} | \Theta) | \mathbf{Y}, \hat{\Theta}^{(t)} \right] \equiv \log f(\mathbf{Y}, \bar{\mathbf{Z}}^{(t)} | \Theta),$$

onde a igualdade é justificada pelo facto de $\log p(\mathbf{Y}, \mathbf{Z} | \Theta)$ ser uma função linear de \mathbf{Z} e onde cada elemento $\bar{z}_{ik}^{(t)}$ de $\bar{\mathbf{Z}}^{(t)}$ é dado por

$$\bar{z}_{ik}^{(t)} = E \left[Z_{ik} | \mathbf{Y}, \hat{\Theta}^{(t)} \right] = P \left[Z_{ik} = 1 | \underline{y}_i, \hat{\Theta}^{(t)} \right] = \frac{\alpha_k f(\underline{y}_i | \underline{\theta}_k^{(t)})}{\sum_{k=1}^K \alpha_k f(\underline{y}_i | \underline{\theta}_k^{(t)})},$$

e t indica a iteração que está a ser executada.

Passo M: Calcula as estimativas dos parâmetros mediante a maximização do valor esperado do logaritmo da verosimilhança completa obtida no passo E

$$\hat{\Theta}^{(t+1)} = \arg \max_{\Theta} \log p(\mathbf{Y}, \bar{\mathbf{Z}}^{(t)} | \Theta) + \log p(\Theta), \quad (1)$$

onde $p(\Theta)$ é a probabilidade *a priori* considerada quando se pretende obter estimativas MAP; quando se pretende encontrar os estimadores de ML, a parcela $\log p(\Theta)$ é omitida.

3 O algoritmo proposto: EM-MML-mix

Para estimar os parâmetros da mistura de multinomiais e gaussianas e simultaneamente o número de componentes, propomos uma variante do algoritmo EM, designado EM-MML-mix. Tomámos por base dois trabalhos [11, 18] que usam um critério MML e desenvolveram uma variante do algoritmo EM para estimação de mistura de gaussianas e multinomiais, respectivamente.

O critério MML privilegia um modelo estatístico que descreva os dados de forma sucinta, no sentido da teoria da informação. Assim,

para uma v.a. Y com f.(d.)p. $p(y|\Theta)$, o comprimento de codificação óptimo (em bits) de uma observação y que é dado por $l(y, \theta) = -\log_2 p(y|\Theta)$ (eventualmente adicionada de $\log_2 p(\Theta)$ quando os parâmetros Θ são desconhecidos) deverá ser o menor possível. Para misturas, esta função (cujo desenvolvimento encontra-se em [3]) é

$$l(y, \Theta) = -\log p(\Theta) - \log p(y|\Theta) + \frac{1}{2} \log |I(\Theta)| + \frac{(K-1)KN}{2} (1 - \log(12)),$$

onde $|I(\Theta)|$ é o determinante do valor esperado da matriz de Fisher, $I(\Theta) \equiv -E \left[\frac{\partial^2}{\partial \theta^2} \log p(Y|\Theta) \right]$, e N é o número de parâmetros a ser estimado em cada componente de mistura. No contexto de agrupamento com modelos de mistura, para ultrapassar alguns problemas de cálculo, em [11] o valor esperado da matriz de Fisher foi calculado considerando os dados completos, ou seja, $I_c(\Theta) \equiv -E \left[\frac{\partial^2}{\partial \theta^2} \log p(Y, Z|\Theta) \right]$. Os autores também usaram *priors* independentes de Jeffreys para os parâmetros de mistura, chegando assim à seguinte função *message length*

$$l(y, \Theta) = \frac{N}{2} \sum_{k: \alpha_k > 0} \log \left(\frac{n \alpha_k}{12} \right) + \frac{k_{nz}}{2} \log \frac{n}{12} + \frac{k_{nz}(N+1)}{2} - \log p(y, \Theta)$$

onde k_{nz} o número de componentes com probabilidade diferente de zero.

Para estimar os parâmetros do modelo de mistura de multinomiais e gaussianas, propomos uma variante do algoritmo EM. Este algoritmo, EM-MML-mix permite estimar, simultaneamente, o número de segmentos e os parâmetros distribucionais associados às variáveis base de segmentação.

Algoritmo 3.1

Passo E: O passo E é igual ao do algoritmo EM

$$\hat{z}_{ik}^{(t)} = \frac{\alpha_k f(\underline{y}_i | \underline{\theta}_k^{(t)})}{\sum_{j=1}^K \alpha_j f(\underline{y}_i | \underline{\theta}_j^{(t)})},$$

para $i = 1, \dots, n$ e $k = 1, \dots, K$;

onde $f(\underline{y}_i | \underline{\theta}_k^{(t)}) = \prod_{m=1}^M f(\underline{y}_{im} | \underline{\theta}_{km}^{(t)}) \prod_{g=1}^G f(\underline{y}_{ig} | \underline{\theta}_{kg}^{(t)})$

Passo M: Atualiza as estimativas dos parâmetros do modelo de mistura:

- as probabilidades de mistura

$$\hat{\alpha}_k^{(t+1)} = \frac{\max \left\{ 0, \sum_{i=1}^n \bar{z}_{ik}^{(t)} - \frac{N}{2} \right\}}{\sum_{j=1}^K \max \left\{ 0, \sum_{i=1}^n \bar{z}_{ij}^{(t)} - \frac{N}{2} \right\}},$$

para $k = 1, \dots, K$ e onde N é o número de parâmetros a estimar em cada componente de mistura.

Repare-se que no caso de algum valor $\hat{\alpha}_k^{(t+1)}$ ser zero a k -ésima componente da mistura é eliminada. Os parâmetros das componentes da mistura com $\hat{\alpha}_k^{(t+1)} = 0$ não precisam ser calculados uma vez que não contribuem para a verosimilhança, pelo que depois de calculados os valores de $\hat{\alpha}_k^{(t+1)}$ só se calculam os parâmetros das componentes cuja probabilidade de mistura é diferente de zero, $\hat{\alpha}_k^{(t+1)} > 0$.

- os parâmetros da multinomial

$$\hat{\theta}_{kmc}^{(t+1)} = \frac{\sum_{i=1}^n \bar{z}_{ik}^{(t)} y_{imc}}{n_m \sum_{i=1}^n \bar{z}_{ik}^{(t)}},$$

para $k = 1, \dots, K, m = 1, \dots, M$ e $c = 1, \dots, C_m$.

- os parâmetros da gaussiana

$$\hat{\mu}_{kg}^{(t+1)} = \frac{\sum_{i=1}^n \tilde{z}_{ik}^{(t)} y_{ig}}{\sum_{i=1}^n \tilde{z}_{ik}^{(t)}}$$

$$\hat{\sigma}_{kg}^{(t+1)} = \frac{\sum_{i=1}^n \tilde{z}_{ik}^{(t)} (y_{ig} - \hat{\mu}_{kg}^{(t+1)})^2}{\sum_{i=1}^n \tilde{z}_{ik}^{(t)}}$$

para $k = 1, \dots, K$ e $g = 1, \dots, G$.

4 Agrupamento usando o EM-MML-mix

4.1 Agrupamento de dados sintéticos

Para avaliar o desempenho do algoritmo, começamos por aplicá-lo a dados sintéticos. Consideramos 2 conjuntos de dados gerados a partir de duas componentes de mistura. Num dos conjuntos consideramos apenas uma variável métrica e uma variável categorial e no outro consideramos mais uma variável categorial. Para estes 2 conjuntos, geraram-se dados de dimensões diferentes (250 e 1000) e em ambos os casos consideraram-se componentes de mistura equilibradas (120 vs 130 e 450 vs 550) e componentes de mistura não equilibradas (50 vs 200 e 800 vs 200), perfazendo um total de 8 amostras de dados sintéticos. Em cada um dos casos geraram-se 10 réplicas, tendo-se corrido o EM-MML-mix e escolhido a solução que apresentava o menor message length.

O algoritmo EM-MML-mix recupera os dois segmentos. No caso da ou das variáveis categoriais, as probabilidades associadas a cada uma das categorias são exatamente iguais, se arredondadas a uma

casa decimal. Quanto à variável métrica, os resultados não são tão bons, uma vez que se obtêm estimativas próximas para as médias, sendo as dos desvios padrão superiores aos valores originais.

4.2 Agrupamento das regiões do European Social Survey

Dados do ESS

Os dados reais analisados são provenientes do European Social Survey (ESS). Este é um inquérito transnacional dirigido aos cidadãos europeus que se realiza de dois em dois anos, desde 2001. O objetivo da análise é agrupar as 250 regiões (de 21 países) do European Social Survey (round 7, 2014), atendendo a indicadores relacionados com o trabalho, nomeadamente uma variável binária Y_1 – *É responsável por supervisionar outros no trabalho?* e uma métrica Y_2 – *Número de horas contratadas por semana no trabalho*. Os dados referidos às regiões são obtidos mediante soma ponderada de respostas "sim" e "não" à questão Y_1 (codificadas com 1 e 0 respetivamente) e mediante média ponderada referida ao número de horas contratadas. A ponderação atende à dimensão da população e ao peso pós-estratificação. Sendo assim trabalha-se com as variáveis *Número de supervisores* e *Número médio de horas de trabalho*, por região.

Resultados Comparativos

Para uma análise comparativa dos resultados da variante EM-MML-mix é usada, como alternativa, a tradicional metodologia de estimação baseada no critério da máxima verosimilhança (ML) seguida de critérios de teoria da informação para a determinação do número de segmentos. São usados os critérios BIC, ICL, AIC e variantes. Estes critérios adicionam à função de verosimilhança (que se pretende maximizar) uma penalização da complexidade do modelo que dependerá da dimensão da amostra e/ou do número de parâmetros a estimar (favorecendo um modelo mais parcimonioso). Em resul-

tado da análise são obtidos dois segmentos, independentemente do critério adotado. Há, no entanto, diferenças entre os dois agrupamentos e constata-se que os indicadores de coesão-separação utilizados – índice Silhueta [14] e Calinski e Harabasz [7] – apontam para uma melhor solução produzida pela estimação de modelo de mistura usando o EM tradicional (Tabela 1). O tempo de computação favorece claramente a metodologia que se propõe.

Tabela 1: Qualidade dos agrupamentos obtidos e tempo de computação associado

Critério	BIC, AIC, CAIC, AIC3, ICL	EM-MML
Número de grupos	2	2
Índice Silhueta	0.541	0.53
Calinski e Harabasz	339.613	323.493
Tempo de computação	26,521 s	6,278 s

Os segmentos obtidos

Em resultado da análise efetuada e atendendo aos indicadores de qualidade de agrupamento, as regiões agrupam-se nos dois segmentos propostos pela aplicação da metodologia de estimação por ML seguida de critérios habituais da teoria de informação para a determinação do número de grupos. No primeiro grupo encontram-se 62 regiões e no segundo 188. Na Figura 1 representa-se a localização geográfica das regiões dos dois segmentos.

De acordo com o seu perfil, o segmento 1 caracteriza-se por ter, em média, mais trabalhadores em funções de supervisão (37%) e por uma média de horas de trabalho que ronda as 32h semanais. No segmento 2 a média de horas de trabalho sobe para 39h e apenas 26% são responsáveis por supervisionar outros no trabalho.

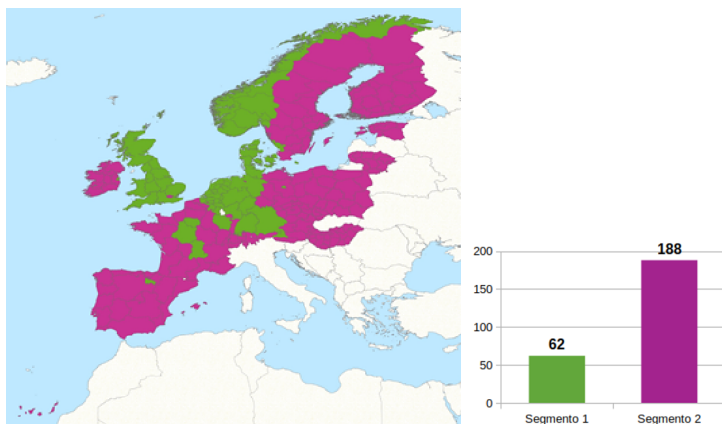


Figura 1: Distribuição geográfica dos segmentos constituídos

5 Conclusões

Neste trabalho, propusemos uma variante do algoritmo EM, o EM-MML-mix, para agrupar dados agregados mistos. O algoritmo proposto permite estimar simultaneamente os parâmetros de uma mistura finita de multinomiais e gaussianas, assim como o número de componentes da mistura (número de segmentos), com base no critério *Minimum Message Length*. Quando lidamos apenas com dados categoriais o EM-MML apresenta resultados melhores que o ICL e semelhantes aos obtidos usando BIC, AIC, CAIC e AIC3 [18]. Na presença de dados mistos e no conjunto limitado de testes por agora efetuados sobre dados gerados e um conjunto de dados reais, a metodologia proposta destaca-se somente pelo seu reduzido tempo de computação. Esta será uma vantagem relevante ao trabalhar com um grande volume de dados. Sobre os dados sintéticos, observa-se, contudo, alguma imprecisão na recuperação da estrutura de dados gerados. Por isso, em trabalhos futuros, o EM-MML-mix deverá ser

melhorado e testado em conjuntos de dados com mais variáveis e de maior dimensão, de forma a reavaliar esta vantagem e a obter uma melhor compreensão do seu desempenho.

Referências

- [1] Ahmad, A., Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, 63(2), 503–527.
- [2] Akaike, H.(1973). Maximum Likelihood Identification of Gaussian Autorregressive Moving Average Models. *Biometrika*, 60, 255–265.
- [3] Baxter, R. A. e Olivier, J. J. (2000). Finding overlapping components with MML. *Statistics and Computing*, 10(1), 5–16.
- [4] Biernacki, C., Celeux, G., Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern analysis and Machine Intelligence*, 22, 719–25.
- [5] Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345–370.
- [6] Bozdogan, H. (1994). Mixture-Model Cluster Analysis using Model Selection criteria and a new Informational Measure of Complexity. *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Approach*, 69–113.
- [7] Calinski, R. B., Harabasz, J. (1974) A dendrit method for cluster analysis. *Communications in Statistics*, 3(1), 1–27.
- [8] Chiu, T., Fang, D., Chen, J., Wang, Y., Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In Provost, R., Srikant, R., (eds.): *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 263–268.
- [9] Dempster, A., Laird, N., Rubin, D. (1997). Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. *Journal of Royal Statistical Society*, 39, 1–38, Series B.

- [10] Everitt, B. S. (1988). A finite mixture model for the clustering of mixed-mode data. *Statistics and probability letters*, 6(5), 305–309.
- [11] Figueiredo, M. A. T., Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3), 381–396.
- [12] Fonseca, J. R., Cardoso, M. G. (2007). Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis*, 11(2), 155–173.
- [13] Hunt, L., Jorgensen, M. (1999). Theory and Methods: Mixture model clustering using the MULTIMIX program. *Australian and New Zealand Journal of Statistics*, 41(2), 154–171.
- [14] Kaufman, L., Rousseeuw, P. J. (1990). *Finding groups in data: an Introduction to cluster analysis*. Wiley, NY.
- [15] Marbac, M., Sedki, M. (2017). Model-based clustering of Gaussian copulas for mixed data. *Communications in Statistics - Theory and Methods*, 43(26), 11635–11656.
- [16] Saha, A., Das, S. (2018). Clustering of fuzzy data and simultaneous feature selection: A model selection approach. *Fuzzy Sets and Systems*, 340, 1–37.
- [17] Schwarz, G. (1978). Estimating the Dimension of a Model *The Annals of Statistics*, 6, 461–464.
- [18] Silvestre, C. (2015). *Clustering with Discrete Mixture Models - An integrated approach for model selection*. Tese de Doutoramento. ISCTE - IUL.
- [19] Silvestre, C., Cardoso, M. G. M. S. and Figueiredo, M. (2015). Feature selection for clustering categorical data with an embedded modelling approach. *Expert Systems*, 32, 444–453.
- [20] Vermunt, J., Magidson, J. (2002). *Applied latent class analysis*. JA Hagenaaers and AL McCutcheon, Cambridge: Cambridge University Press.
- [21] Wallace, C. S., Boulton, D. M. (1968). An information measure for classification. *The Computer Journal*, 11(2), 195–209.
- [22] Wedel, M., Kamakura, W. (2002). *Market Segmentation- Conceptual and Methodological Foundations*. Vol. 8. Edições Springer Science & Business Media, New York.

Método das maiores observações anuais: Aplicação ao triplo-salto masculino

Domingos Silva

Universidade de Évora, Centro de Investigação em Matemática e Aplicações, Portugal, *domingosjlsilva@gmail.com*

Frederico Caeiro

Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia & Centro de Matemática e Aplicações, Portugal, *fac@fct.unl.pt*

Manuela Oliveira

Universidade de Évora, Escola de Ciências e Tecnologia, Departamento de Matemática & Centro de Investigação em Matemática e Aplicações, Portugal, *mmo@uevora.pt*

Palavras-chave: Método das r maiores observações; Teoria de valores extremos; Triplo salto.

Resumo: Na teoria de valores extremos, o método das r maiores observações em cada bloco (r -MO) é uma extensão do método dos máximos de blocos ou método de Gumbel. Nesta abordagem usamos as maiores r observações de cada bloco para estimar os parâmetros do modelo de valores extremos multivariado. Neste trabalho consideramos os melhores resultados do triplo-salto do atletismo masculino, entre 1980 e 2016, e usamos o método r -MO anuais ($r = 1, \dots, 10$) para estimar diversos parâmetros extremais. Os resultados apontam para a existência de um parâmetro de forma negativo, representativo da cauda direita leve e com limite superior do suporte finito, independentemente do valor de r . A estimativa do limite superior do suporte varia entre 18.43m ($r = 4$) e 18.57m ($r = 1$). A probabilidade de obter um novo máximo mundial varia entre 0.297% ($r = 4$) e 0.717% ($r = 1$). Com base no intervalo com 95% de confiança, o nível de retorno a 50 anos contém um novo recorde do mundo.

1 Introdução

O triplo-salto é uma disciplina do atletismo, similar ao salto em comprimento, podendo ser praticada *indoor* e *outdoor*. São fatores importantes para um bom triplo-salto a velocidade de deslocamento, a posição do corpo na chamada (i.e., deslocamento máximo horizontal do centro de gravidade no momento da saída), a velocidade da impulsão, o ângulo da impulsão, a altura da impulsão, a resistência ao ar, a posição do corpo na receção ao solo e ações posteriores (Hay [5]). Ao ar livre, o primeiro registo que se conhece nos homens data de 1826 e pertence a Andrew Beatti (GBI, Great Britain & Ireland), com a marca de 12.95m. A IAAF - International Amateur Athletic Federation passou a homologar as marcas a partir de 1912 inclusive (Hymans e Matrahazi [6]). Na Tabela 1 apresentamos alguns recordes do triplo-salto masculino até 31-12-2016. Desde a vitória de Nelson Évora nos Campeonatos do Mundo

Tabela 1: Alguns recordes do triplo-salto masculino até 31-12-2016.

Atleta	Marca	Ano	Observação
Andrew Beattie (GBI)	12.95m	1826	1º recorde do mundo
Daniel Ahearn (GBI)	15.52m	1912	1º recorde do mundo da IAAF
Kenny Harrison (EUA)	18.09m	1996	recorde olímpico (Atlanta)
Jonathan Edwards (ING)	18.29m	1995	recorde do mundo
Nelson Évora (POR)	17.74m	2007	recorde nacional

de 2007, em Osaka (Japão), com a marca de 17.74m que Portugal tem vindo a ganhar tradição no triplo-salto. Porém, até 31-12-2016, o recorde nacional de Nelson Évora ocupava a 66ª melhor marca de sempre. Esta marca só seria recorde do mundo até 15-10-1975. Em termos individuais, Nelson Évora é, até à data deste estudo, o 25º melhor saltador do triplo-salto de sempre. Para ajudar a compreender o fenómeno dos recordes, recorreremos à Teoria de Valores Extremos (EVT, *Extreme Value Theory*), dado possuir modelos usados na inferência de acontecimentos atípicos que são mais extremos do que quaisquer outros já observados. O teorema dos tipos extre-

mais de Fisher, Tippet e Gnedenko desempenha um papel crucial. Este teorema garante que se o máximo amostral, linearmente normalizado, convergir para uma distribuição não degenerada, então essa distribuição é a distribuição Generalizada de Valores Extremos (GEV, *Generalized Extreme Value*). Neste trabalho baseamos a inferência no método paramétrico das r -maiores observações de blocos (r -MO), também designado GEV-multivariado ou GEV-processo extremal (Smith [8], Tawn [10]). É usual considerarmos que cada bloco é constituído por todas as observações registadas durante um ano. Weissman [11] foi quem primeiro abordou o modelo das r -MO. Posteriormente, Smith [8] desenvolveu a metodologia estatística tal como hoje a conhecemos. Comparativamente ao método dos máximos de blocos, a abordagem r -MO, com $r > 1$ usa mais informação proveniente da amostra. Esta abordagem baseia-se na distribuição assintótica conjunta das r -MO em cada bloco (e.g., ano), com $r > 1$, as quais generalizam a distribuição GEV. Na prática, não é fácil a escolha de r . As dificuldades são análogas à escolha do limiar no método *Peaks Over Threshold* (POT). Um r muito pequeno é suscetível de originar aumento da variância dos estimadores, mas se r for muito grande pode ocorrer viés (Coles [2]). Assim, na presença não apenas do valor máximo de cada ano mas das dez maiores observações de cada ano relativas ao triplo-salto masculino no período de 1980 a 2016, iremos aplicar a metodologia das r -MO para obter estimativas de algumas quantidades de interesse, nomeadamente quantis extremas, a probabilidade de excedência, o limite superior do suporte e os valores de retorno.

2 Metodologia

2.1 Resultados Fundamentais em EVT

Seja (X_1, \dots, X_n) uma amostra aleatória (a.a.), i.e., uma sequência de n variáveis aleatórias (v.a.'s) independentes e identicamente distribuídas (i.i.d.'s), com função de distribuição (f.d.) comum $F(\cdot)$ e

$M_n = \max(X_1, \dots, X_n)$ o máximo da amostra. Então $M_n \xrightarrow{p} x^F$, com $x^F := \sup\{x \in \mathbb{R} : F(x) < 1\}$ o limite superior do suporte de F . Contudo, sendo F desconhecida, tal compromete qualquer inferência para o máximo, onde é necessário que este tenha um comportamento assintótico não-degenerado. Assim, à semelhança do que acontece com o Teorema do Limite Central para a soma de v.a.'s, a normalização do máximo faz-se recorrendo a sucessões de constantes reais $a_n > 0$ e b_n ($n \in \mathbb{N}$), tais que:

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow[n \rightarrow \infty]{d} G(x), \quad (1)$$

para uma f.d. G não-degenerada, pelo que G pertence à família GEV (Gnedenko [4]),

$$GEV(x|\lambda, \delta, \gamma) = \begin{cases} \exp\left(-(1 + \gamma x)^{-\frac{1}{\gamma}}\right), & 1 + \gamma x > 0, \text{ se } \gamma \neq 0 \\ \exp(-\exp(-x)), & x \in \mathbb{R}, \text{ se } \gamma = 0, \end{cases} \quad (2)$$

com $x = (y - \lambda)/\delta$, $\lambda \in \mathbb{R}$, $\delta > 0$ e $\gamma \in \mathbb{R}$ os parâmetros de localização, escala e forma, respetivamente. Nas aplicações com dados reais, as constantes a_n e b_n são desconhecidas e podem ser incorporadas nos parâmetros de escala e localização da distribuição limite em (2). O parâmetro de forma γ ou índice de valores extremos é fulcral, pois determina o comportamento da cauda direita da distribuição. Se $\gamma = 0$, F tem cauda direita do tipo exponencial e com limite superior do suporte de F finito ($x^F < \infty$) ou infinito ($x^F = \infty$); se $\gamma > 0$, F tem cauda direita pesada do tipo polinomial negativo e com limite superior do suporte de F infinito ($x^F = \infty$); se $\gamma < 0$, F tem cauda direita leve e com limite superior do suporte de F finito ($x^F < \infty$). No método dos blocos, a amostra é dividida em m subamostras de igual dimensão. O modelo GEV em (2) é depois usado para modelar a amostra constituída pelo valor máximo de cada bloco. Uma das dificuldades da utilização do modelo (2) resulta da escassa informação usada e consequente elevada variância dos estimadores.

Para usarmos de modo mais eficiente a informação da amostra, vamos também considerar o modelo para as r maiores estatísticas ordinais, $X_i^{(1)} \geq \dots \geq X_i^{(r)}$ ($M_{n;i} = X_i^{(1)}$). Vamos assumir que existem sucessões de constantes normalizadoras $a_n > 0$ e b_n ($n \in \mathbb{N}$), tais que $(M_n - b_n)/a_n$ converge em distribuição para o modelo GEV em (2). Então, a distribuição limite conjunta das r maiores estatísticas ordinais normalizadas

$$\left(\frac{X^{(1)} - b_n}{a_n}, \frac{X^{(2)} - b_n}{a_n}, \dots, \frac{X^{(r)} - b_n}{a_n} \right) \quad (3)$$

é a família de modelos GEV_r (Dwass [3]) com função densidade conjunta

$$g^{(r)}(x^{(1)}, \dots, x^{(r)} | \lambda, \delta, \gamma) = G(x^{(r)} | \lambda, \delta, \gamma) \times \prod_{j=1}^r \frac{g(x^{(j)} | \lambda, \delta, \gamma)}{G(x^{(j)} | \lambda, \delta, \gamma)}, \quad x^{(1)} \geq \dots \geq x^{(r)} \quad (4)$$

sendo $g = G'$ a função densidade do modelo GEV. Quando $r = 1$, a distribuição GEV_r em (4) é exatamente a distribuição GEV dada em (2). As constantes a_n e b_n são usualmente desconhecidas e habitualmente incorporadas nos parâmetros de escala e de localização da distribuição limite.

Observação 2.1 *A distribuição limite do máximo em (2) e a distribuição limite das r maiores estatísticas ordinais em (4) partilham os mesmos parâmetros de localização, escala e forma $(\lambda, \delta, \gamma)$.*

Observação 2.2 *A escolha de r deve ser feita de modo prudente, uma vez que temos a habitual troca entre variância e viés.*

Reduzidos níveis de r geram poucos dados, implicando uma variância elevada, e grandes níveis de r provavelmente originam viés (Coles, [2]). Em geral, r necessita de ser pequeno em relação ao tamanho do bloco (não ao número de blocos), uma vez que à medida que r aumenta, diminui a taxa de convergência da distribuição das r maiores estatísticas ordinais, para a sua distribuição limite conjunta (Smith [8]).

2.2 Estimação dos parâmetros do modelo GEV_r

Considere a amostra disposta em m -blocos, cada um com dimensão n . Em cada bloco extraem-se as r maiores observações, obtendo-se assim um conjunto de m vetores aleatórios r -dimensionais, $x_i^{(1)}, \dots, x_i^{(r)}$, $i = 1, \dots, m$. A estimação é usualmente feita pelo método da máxima verosimilhança (MV). A função de verosimilhança $L(\lambda, \delta, \gamma)$ é obtida a partir de (4) e dada por

$$L_r(\lambda, \delta, \gamma) = \prod_{i=1}^m g^{(r)}(x_i^{(1)}, \dots, x_i^{(r)} | \lambda, \delta, \gamma). \quad (5)$$

Para facilitar a manipulação matemática da verosimilhança, utiliza-se frequentemente a função de log-verosimilhança,

$$\ell_r(\lambda, \delta, \gamma) = \log L_r(\lambda, \delta, \gamma). \quad (6)$$

A maximização da função de log-verosimilhança não tem solução analítica. Portanto, para se obter as estimativas de MV dos parâmetros $(\lambda, \delta, \gamma)$ do modelo GEV_r é necessário usar técnicas iterativas de otimização numérica. Smith [7] mostrou que as propriedades assintóticas dos estimadores de MV dependem do parâmetro de forma γ . Quando $\gamma > -0.5$, verificam-se as condições de regularidade e os estimadores de MV têm as propriedades assintóticas usuais: consistência, eficiência, invariância e normalidade; quando $-1 < \gamma < -0.5$ os estimadores de MV são geralmente obtidos mas não têm as propriedades usuais; quando $\gamma < -1$, os estimadores de MV não existem ou são inconsistentes.

Os erros-padrão (*se*) para $(\lambda, \delta, \gamma)$, são calculados usando a matriz de informação de Fisher. Detalhes sobre este procedimento podem ser observados em Coles ([2], p.32). Se θ representar uma das componentes do vetor $(\lambda, \delta, \gamma)$, o intervalo com nível de confiança $100(1 - \alpha)\%$, é dado por,

$$IC_{100(1-\alpha)\%}(\theta) = \hat{\theta} \pm z_{1-\alpha/2} se(\hat{\theta}), \quad (7)$$

onde $z_{1-\alpha/2}$ é o quantil $(1 - \alpha/2)$ da distribuição normal padrão.

2.3 Validação do modelo GEV_r e escolha de r

Para verificação do ajustamento do modelo, Smith [8] e Tawn [10] usam o gráfico de probabilidade (PP *plot*) para a distribuição marginal da k -ésima estatística ordinal de topo ($1 \leq k \leq r$). Para além do PP *plot*, também o gráfico de quantis (QQ *plot*) é sugerido por Coles [2]. Por sua vez, Soares and Scotto [9] sugerem o uso do teste da razão de verosimilhanças para a escolha do valor de r . O teste baseia-se na estatística *Deviance* para testar a hipótese nula $r = i$ contra a hipótese alternativa $r = i + 1$,

$$D(i) = 2\{\ell_{i+1}(\lambda, \delta, \gamma) - \ell_i(\lambda, \delta, \gamma)\} \sim \chi_1^2, \quad i = 1, 2, \dots \quad (8)$$

com $\ell_r(\lambda, \delta, \gamma)$ definida em (6). A um nível de significância α , rejeita-se a hipótese nula se $D(i) \geq \chi_{1, 1-\alpha}^2$. Para o nível de significância $\alpha = 0.05$, $\chi_{1, 0.95}^2 = 3.841$.

Smith [8] e An and Pandley [1] sugerem que se observem os erros-padrão das estimativas dos parâmetros $(\lambda, \delta, \gamma)$ nos diferentes níveis de r , indicando que quanto mais reduzidos, melhor será a qualidade do modelo.

2.4 Estimação de outros parâmetros de interesse

Nas aplicações com dados reais a estimação de outros parâmetros que dependem de λ , δ e γ é extremamente importante. Referimos a seguir os parâmetros mais importantes:

- Probabilidade de excedência: é a probabilidade de ultrapassar um nível x elevado. Pode ser estimada através de

$$P[X > x] := 1 - G(x|\hat{\lambda}, \hat{\delta}, \hat{\gamma}), \quad (9)$$

com G definida em (2).

- Quantil extremal: o quantil extremal de probabilidade $1 - p$ (também chamado de quantil extremal de probabilidade de excedência p), denotado por q_{1-p} , é o valor que é excedido

com probabilidade p ($p < 1/n$). Os quantis extremais obtêm-se invertendo a f.d. GEV em (2). A sua estimativa é

$$\hat{q}_{1-p} := G^{\leftarrow}(1-p|\hat{\lambda}, \hat{\delta}, \hat{\gamma}). \quad (10)$$

- **Nível de retorno:** o nível de retorno para t -anos, $U(t)$, é o nível que é excedido em média uma vez em cada t -anos. Especificamente, o nível de retorno para t -anos é o quantil extremal de probabilidade $1 - 1/t$ da distribuição GEV.
- **Período de retorno:** o período de retorno T de um nível elevado é o tempo médio de espera até à ocorrência de um evento de magnitude superior ao evento extremo de nível x . Pode ser estimado por:

$$\hat{T} = \frac{1}{1 - G(x|\hat{\lambda}, \hat{\delta}, \hat{\gamma})}. \quad (11)$$

- **Limite superior do suporte** (para $\gamma < 0$) Se $\hat{\gamma} < 0$, a estimativa do limite superior do suporte, $x^F = \sup\{x \in \mathbb{R} : F(x) < 1\}$, é finita e é dada por $\hat{x}^F := \hat{q}_0 = \hat{U}(\infty) = \hat{\lambda} - \frac{\hat{\delta}}{\hat{\gamma}}$.

Observação 2.3 *Para os parâmetros anteriormente apresentados podem ser construídos intervalos de confiança. O intervalo é obtido usando a matrix de variâncias e covariâncias e o método delta.*

3 Aplicação ao triplo-salto masculino

Os dados referentes ao triplo-salto masculino foram obtidos no Website <http://www.all-athletics.com/en-us/all-time-lists>, no período de 1980 a 2016. Neste estudo, cada bloco corresponde a um ano. Obtivemos, para cada ano, os 10 melhores resultados homologados pela IAAF. Utilizamos o método das r -MO anuais para estimar os parâmetros da distribuição GEV_r em (4), considerando todos os possíveis valores de r ($1 \leq r \leq 10$). A Tabela 2 apresenta o valor da log-verosimilhança maximizada e as estimativas dos parâmetros de

localização (λ), escala (δ) e forma (γ) com os respectivos erros-padrão (se) e os intervalos com 95% de confiança ($IC95\%$).

Tabela 2: Log-verosimilhança maximizada e estimativas dos parâmetros de localização, escala e forma, com os respectivos erros-padrão e os intervalos com 95% de confiança, do modelo r -MO.

r	ℓ	$\hat{\lambda}$	$se(\hat{\lambda})$	$\hat{\delta}$	$se(\hat{\delta})$	$\hat{\gamma}$	$se(\hat{\gamma})$
1	-8.93	17.339408 (17.67; 17.80)	0.033	0.1851143 (0.141; 0.230)	0.023	-0.2210754 (-0.400; -0.040)	0.092
2	-52.33	17.683949 (17.72; 17.82)	0.027	0.1785661 (0.151; 0.206)	0.014	-0.2526322 (-0.370; -0.130)	0.061
3	-108.53	17.688654 (17.72; 17.82)	0.025	0.1728959 (0.151; 0.195)	0.011	-0.2407046 (-0.340; -0.140)	0.052
4	-176.33	17.830809 (17.74; 17.83)	0.023	0.1701999 (0.152; 0.188)	0.009	-0.2630911 (-0.340; -0.180)	0.041
5	-255.21	17.824574 (17.74; 17.83)	0.022	0.1648447 (0.148; 0.182)	0.009	-0.2477381 (-0.320; -0.170)	0.038
6	-335.32	17.799112 (17.74; 17.82)	0.022	0.1633397 (0.147; 0.180)	0.008	-0.2390355 (-0.310; -0.170)	0.036
7	-408.90	17.776463 (17.73; 17.82)	0.022	0.1662146 (0.149; 0.183)	0.009	-0.2402128 (-0.310; -0.170)	0.037
8	-493.95	17.764347 (17.73; 17.82)	0.022	0.1666516 (0.150; 0.184)	0.009	-0.2375130 (-0.310; -0.170)	0.036
9	-589.77	17.748556 (17.73; 17.82)	0.021	0.1655585 (0.149; 0.183)	0.009	-0.2307629 (-0.300; -0.160)	0.035
10	-681.85	17.781159 (17.74; 17.82)	0.021	0.1649249 (0.149; 0.181)	0.008	-0.2398879 (-0.300; -0.180)	0.031

Observa-se que valores crescentes de r correspondem a:

- Estabilidade das estimativas ($\hat{\lambda}, \hat{\delta}, \hat{\gamma}$). A estabilidade é mais evidente entre os modelos $r=2, \dots, 10$ para $\hat{\lambda}$, entre $r=5, \dots, 10$ para $\hat{\delta}$ e entre $r=2, 3$ bem como $r=5, \dots, 10$ para $\hat{\gamma}$.
- Decréscimo dos erros-padrão das estimativas dos parâmetros. O modelo GEV_r , com $r=1$, tem os mais elevados erros-padrão, e o modelo GEV_r , com $r=10$, os mais baixos, o que sugere a existência de um melhor ajustamento do modelo quando se consideram mais observações por ano, para além do máximo anual.
- Em todos os valores de r considerados, os intervalos com 95% de confiança para γ só contêm valores negativos, o que nos leva a rejeitar a hipótese do modelo ser Gumbel. Assim, a distribuição Weibull parece ser uma escolha razoável para o

conjunto de dados referentes ao triplo-salto masculino. Neste caso, a distribuição tem cauda superior leve e limite superior do suporte finito ($x^F < \infty$).

A Tabela 3 apresenta as estimativas do limite superior do suporte de F , da probabilidade de excedência do nível $x = 18.29\text{m}$ (atual recorde do mundo), do período de retorno do nível $x = 18.29$ e dos quantis extremos a níveis de probabilidade (de excedência) de 0.5%, 0.25% e 0.1%, segundo o método das r -MO anuais, para $r=1, \dots, 10$. A estimativa de x^F é mais alta em $r=1$ e mais baixa em $r=4$; nos restantes casos de r , verifica-se uma certa estabilidade. Ou seja, o modelo GEV indica que o recorde do mundo nunca será superior a 18.57m, enquanto que no modelo GEV_4 nunca ultrapassará os 18.43m. A probabilidade do melhor registo anual bater o atual recorde do mundo é de cerca de 0.72% no modelo GEV e inferior a 0.5% nos modelos GEV_r . Os quantis extremos são sempre mais elevados com $r=1$ do que com $r > 1$. Apenas no modelo GEV a estimativa do quantil extremal $\hat{q}_{0.995} = 18.31\text{m}$ ultrapassa o atual recorde do mundo, sendo que nos casos $r = 2$ e $r = 3$ o recorde do mundo é igualado. Noutras estimativas pontuais de quantis extremos, mais concretamente, $\hat{q}_{0.9975}$ e $\hat{q}_{0.999}$, qualquer que seja r , um novo recorde do mundo está sempre presente. Paralelamente, o período de retorno do nível 18.29m é mais baixo com $r=1$. De todos, o modelo GEV_4 , parece ser aquele que considera “mais invulgar” a possibilidade da ocorrência de elevadas performances no triplo-salto masculino. Por exemplo, observando a sua estimativa do período de retorno para o nível 18.29m, espera-se que, em média, uma vez em cada 336.7 anos este nível seja excedido, descendo esse nível para 139.5 anos no caso do modelo GEV.

A Tabela 4 apresenta as estimativas pontuais dos níveis de retorno a 5, 10, 20, 50 e 100 anos, com os respetivos erros-padrão e intervalos com 95% de confiança, ajustadas às performances do triplo-salto masculino. Independentemente do valor de r , verifica-se que o nível de retorno em cada t -anos, apresenta estimativas pontuais, erros-padrão e $IC95\%$ relativamente estáveis. Ou seja, o aumento de

Tabela 3: Estimativas de máxima verosimilhança para o limite superior do suporte, probabilidade de bater o atual recorde do mundo, período de retorno e quantis extremais para 0.5%, 0.25% e 0.1% de probabilidade, do modelo das r -MO ajustado aos dados do triplo-salto masculino para diferentes níveis de r .

r	\hat{x}^F	$P[X > 18.29]$	$T = \frac{1}{\hat{P}(X > 18.29)}$	$\hat{q}_{0.995}$	$\hat{q}_{0.9975}$	$\hat{q}_{0.999}$
1	18.57	0.00717	139.4983	18.31	18.35	18.39
2	18.48	0.00497	201.0745	18.29	18.32	18.35
3	18.49	0.00464	215.6263	18.29	18.32	18.35
4	18.43	0.00297	336.6564	18.27	18.30	18.32
5	18.45	0.00300	333.2350	18.27	18.30	18.33
6	18.46	0.00321	311.8733	18.27	18.30	18.33
7	18.47	0.00364	275.0347	18.28	18.30	18.34
8	18.48	0.00391	263.8084	18.28	18.31	18.34
9	18.49	0.00414	241.7848	18.28	18.31	18.35
10	18.47	0.00338	296.1050	18.27	18.30	18.33

r , não gera modificações substanciais nas estimativas dos níveis de retorno. As estimativas pontuais dos níveis de retorno não preveem um novo recorde do mundo nos próximos 100 anos, mas com base na estimativa do $IC_{95\%}$ para os níveis de retorno a 50 e 100 anos, em todos os níveis de r , observa-se a possibilidade da ocorrência de um novo máximo mundial.

Para a escolha de r , usámos o teste de hipóteses descrito na Secção 2.3. A Tabela 5 apresenta o valor da estatística deviance e o valor de prova. A hipótese nula é sempre rejeitada e concluímos que GEV_r , com $r=10$, gera o melhor modelo.

A Figura 1 apresenta os gráficos de probabilidade (PP-plot) (à esquerda) e os quantis (QQ plot) (à direita), para verificação da qualidade do ajuste do modelo com $r=10$. Os gráficos são referentes à distribuição marginal da k -ésima estatística ordinal superior. Os gráficos de probabilidade e dos quantis parecem indicar um ajustamento razoável, não parecendo colocar em causa a validade do modelo GEV_{10} . Contudo, o modelo com $k = 6$ parece ter o pior ajustamento. De uma forma geral, a nuvem de pontos distribui-se aproximadamente ao longo da reta diagonal $y = x$. Alguns problemas que caracterizam todos os gráficos dos quantis, são a presença de algumas observações na cauda inferior um pouco afastadas da

Tabela 4: Estimativas pontuais dos níveis de retorno a 5, 10, 20, 50 e 100 anos, com os respetivos erros-padrão e intervalo com 95% de confiança entre parêntesis, do modelo das r -MO ajustado aos dados do triplo-salto masculino nos diferentes níveis de r .

r	$\hat{U}(5)$	$\hat{U}(10)$	$\hat{U}(20)$	$\hat{U}(50)$	$\hat{U}(100)$
1	17.97 (0.039) (17.89; 18.05)	18.06 (0.044) (17.98; 18.15)	18.14 (0.052) (18.03; 18.24)	18.22 (0.067) (18.09; 18.35)	18.27 (0.081) (18.11; 18.43)
2	17.99 (0.039) (17.91; 18.07)	18.08 (0.043) (17.99; 18.16)	18.14 (0.050) (18.04; 18.24)	18.21 (0.062) (18.09; 18.33)	18.25 (0.073) (18.11; 18.40)
3	17.99 (0.039) (17.91; 18.06)	18.07 (0.044) (17.98; 18.16)	18.14 (0.050) (18.04; 18.23)	18.21 (0.062) (18.08; 18.33)	18.25 (0.073) (18.11; 18.39)
4	17.99 (0.039) (17.92; 18.07)	18.07 (0.043) (17.99; 18.16)	18.13 (0.049) (18.04; 18.23)	18.20 (0.059) (18.08; 18.31)	18.24 (0.069) (18.10; 18.37)
5	17.99 (0.039) (17.91; 18.07)	18.07 (0.044) (17.98; 18.15)	18.13 (0.050) (18.03; 18.23)	18.20 (0.060) (18.08; 18.31)	18.24 (0.065) (18.11; 18.36)
6	17.99 (0.039) (17.91; 18.06)	18.06 (0.044) (17.98; 18.15)	18.13 (0.045) (18.04; 18.22)	18.19 (0.061) (18.07; 18.31)	18.24 (0.071) (18.10; 18.37)
7	17.99 (0.039) (17.91; 18.06)	18.07 (0.044) (17.98; 18.15)	18.13 (0.046) (18.04; 18.22)	18.20 (0.057) (18.09; 18.31)	18.24 (0.067) (18.11; 18.37)
8	17.99 (0.039) (17.91; 18.06)	18.07 (0.044) (17.98; 18.15)	18.13 (0.050) (18.03; 18.23)	18.20 (0.062) (18.08; 18.32)	18.24 (0.072) (18.10; 18.38)
9	17.99 (0.039) (17.91; 18.06)	18.07 (0.044) (17.99; 18.15)	18.13 (0.050) (18.03; 18.23)	18.20 (0.062) (18.08; 18.32)	18.24 (0.067) (18.11; 18.38)
10	17.99 (0.039) (17.91; 18.06)	18.07 (0.044) (17.98; 18.15)	18.13 (0.049) (18.03; 18.22)	18.20 (0.061) (18.08; 18.32)	18.24 (0.071) (18.10; 18.38)

Tabela 5: Valor da estatística deviance e correspondente valor de prova, entre dois modelos consecutivos.

	$r=1$ vs $r=2$	$r=2$ vs $r=3$	$r=3$ vs $r=4$	$r=4$ vs $r=5$	
D	86.815	112.398	135.587	157.763	
p	<0.001	<0.001	<0.001	<0.001	
	$r=5$ vs $r=6$	$r=6$ vs $r=7$	$r=7$ vs $r=8$	$r=8$ vs $r=9$	$r=9$ vs $r=10$
D	160.234	147.157	170.086	191.649	184.167
p	<0.001	<0.001	<0.001	<0.001	<0.001

reta diagonal, ainda que não seja preocupante dado o interesse deste estudo incidir sobre máximos. Similarmente, na cauda superior para $k = 1, 2, 3$ observa-se a presença de algumas observações ligeiramente mais afastadas.

4 Conclusões

Neste estudo utilizamos o método r -maiores observações de blocos anuais, para estimar os principais parâmetros extremais de interesse no triplo-salto masculino, no período de 1980 a 2016. Qualquer

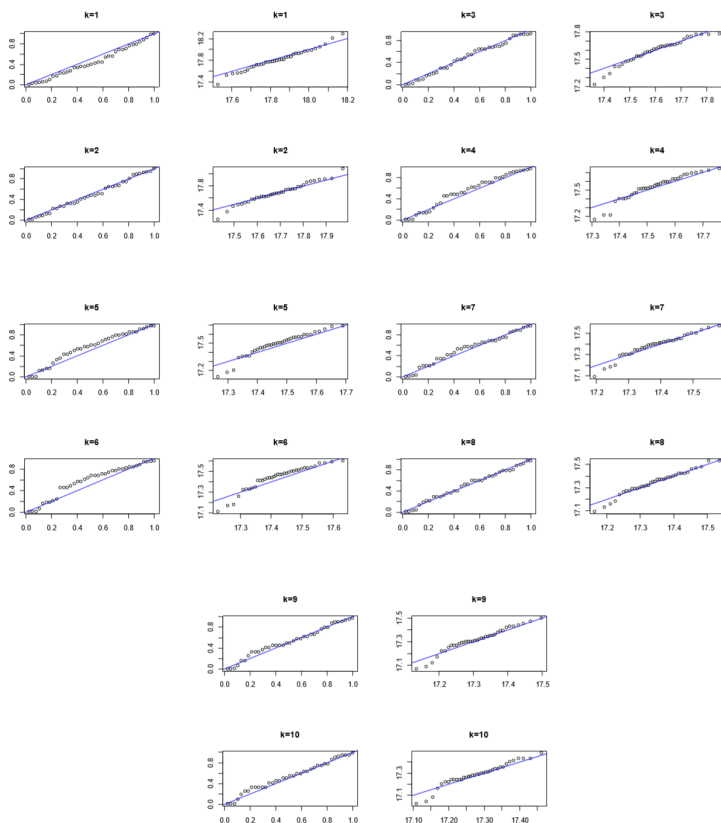


Figura 1: Modelo diagnóstico para os dados do triplo-salto masculino com base no método das maiores observações de blocos anuais, para $r=10$

que seja o número de observações máximas retidas em cada bloco ($r=1, \dots, 10$), a distribuição tem cauda superior leve e limite superior do suporte finito ($\hat{\gamma} < 0$). Verificou-se que a escolha $r = 10$ era

adequada. A probabilidade de bater o atual recorde do mundo é de cerca de 0.72% no modelo GEV e inferior a 0.5% nos restantes modelos (GEV_r , com $r=2, \dots, 10$). Especificamente, no modelo $r=10$, a probabilidade de obter um novo máximo mundial é de cerca de 0.34%, com período de retorno $T=296$ anos. O intervalo com 95% de confiança para os níveis de retorno a 50 e 100 anos, em todos os níveis de r , contém um novo recorde do mundo.

Agradecimentos

Investigação parcialmente suportada pela Fundação para a Ciência e a Tecnologia através do projeto UID/MAT/04674/2019 (CIMA: Centro de Investigação em Matemática e Aplicações) e do projeto UID/MAT/00297/2019 (Centro de Matemática e Aplicações).

Referências

- [1] An, Y., Pandey, M.D. (2007). The r largest order statistics model for extreme wind speed estimation. *Journal of Wind Engineering and Industrial Aerodynamics* 95 (3), 165–182.
- [2] Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. *Springer-Verlag*, London.
- [3] Dwass, M. (1964). Extremal processes. *Ann. Math. Statist.* 35 (4), 1718–1725.
- [4] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d’une série aléatoire. *Ann. Math.* 44 (3), 423–453.
- [5] Hay, J.G. (1993). The Biomechanics of Sports Technique. 4rd ed., *Publisher: Pearson, Wallingford*. United Kingdom.
- [6] Hymans, R., Matrahazi, I., (2015). International Association of Athletics Federations. Progression of IAAF World Records. *Multiprint, Monaco*.
- [7] Smith, R.L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, 72 (1), 67–90.

- [8] Smith, R.L. (1986). Extreme Value Theory Based on the r largest Annual Events. *Journal of Hydrology*, 86 (1–2), 27–43.
- [9] Soares, C.G., Scotto, M.G. (2004). Application of the r largest-order statistics for long-term predictions of significant wave height. *Coastal Engineering* 51 (s5-6), 387–394.
- [10] Tawn, J.A. (1988). Bivariate extreme value theory: model and estimation. *Biometrika* 77 (2), 245–253.
- [11] Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Am. Statist. Assoc.*, 73 (364), 812–815.

Taxas de erros de tipos I e II de procedimentos não paramétricos alternativos à ANOVA com dois fatores para dados discretos

Dulce G. Pereira

CIMA/IIFA e Departamento de Matemática/ECT, Universidade de Évora, *dgsp@uevora.pt*

Anabela Afonso

CIMA/IIFA e Departamento de Matemática/ECT, Universidade de Évora, *aafonso@uevora.pt*

Palavras-chave: Empates; Estatística de Wald; Testes de permutação; Transformação em ordens.

Resumo: Usualmente, nas alternativas à ANOVA paramétrica com dois fatores as observações são substituídas pelas suas ordens. No estudo do desempenho destas alternativas apenas têm sido consideradas distribuições contínuas. Contudo, quando os dados provêm de distribuições discretas propiciam a existência de muitos empates. Neste trabalho, recorrendo a um estudo por simulação, estudamos as taxas de erro de tipo I e II de várias dessas alternativas. Foram considerados delineamentos equilibrados com 2 fatores e dados provenientes de distribuições discretas. Os testes *WTS* e *USP* foram os que mostraram ser liberais. Os testes *L* de Puri & Sen e de *van der Waerden* não mostraram ter um bom desempenho.

1 Introdução

A análise de variância (ANOVA) com dois fatores, A e B , pretende testar se todos os níveis do fator A originam a mesma variância média

na variável resposta (quantitativa), isto é, se possuem um efeito médio igual (analogamente para o fator B), bem como determinar se existe interação entre os dois fatores.

A obtenção de conclusões através da ANOVA deve ser precedida da verificação de algumas condições, sob pena de poder conduzir a inferências erradas. Deve assegurar-se que a sua aplicação só tem lugar quando as observações são independentes e os dados provêm de populações normalmente distribuídas com variância comum.

Em muitas situações não podemos utilizar a ANOVA paramétrica porque os dados não são contínuos e por vezes são ordinais. Nestes casos, podemos recorrer a alternativas não-paramétricas. Muitas destas alternativas podem ser aplicadas quer a dados contínuos quer a dados discretos pois consistem em aplicar transformações aos dados originais (ordens, *scores* normais das ordens, ...) [4]. No entanto, existem poucos estudos sobre o desempenho destas alternativas considerando distribuições discretas. Mansouri *et al.* [6] comparam o desempenho do teste *aligned rank transform* com distribuições do erro contínuas e discretas (Binomial) e não encontraram grandes diferenças. Nos delineamentos 2×2 , Kaptein *et al.* [3] mostraram que, no caso de escalas de Likert, a potência da *ANOVA type statistic* é superior à do teste F da ANOVA.

Neste trabalho, analisamos as probabilidades de erros de tipo I e II das técnicas não paramétricas, considerando um estudo de simulação quando os dados são provenientes de distribuições discretas e delineamentos equilibrados 2×5 , 3×3 e 3×4 .

2 Alternativas não paramétricas

Nos últimos anos foram propostas várias alternativas bastante distintas à ANOVA paramétrica com dois fatores. As técnicas *rank transform* (RT) e *inverse normal transformation* (INT) consistem na substituição das observações pelas suas ordens, ou pelos *scores* normais das ordens, respectivamente, e a posterior aplicação da ANOVA paramétrica usual. A técnica *aligned rank transform* (ART), bem

como a combinação deste método com o *INT* (*ART+INT*), subtrai todos os efeitos que não sejam de primeiro interesse antes de se realizar a ANOVA. A estatística *L* de Puri e Sen (*L de PS*), o teste de van der Waerden (*vdW*), e as *Wald type statistic* (*WTS*) e *ANOVA type statistic* (*ATS*) propõem as suas próprias estatísticas, em alguns casos à custa dos modelos lineares. As alternativas *Wald type statistic permutation* (*WTPS*), *constrained synchronized permutations* (*CSP*) e *unconstrained synchronized permutations* (*USP*) baseiam-se na permutação das observações. A descrição destas técnicas pode ser consultada, por ex., em Hahn *et al.* [2] e Luepsen [4].

Cada uma destas técnicas tem as suas vantagens e desvantagens, não existindo uma que seja melhor do que outra em todos os contextos (Tabela 1). Algumas são muito fáceis de implementar, outras apresentam problemas ao nível do erro de tipo I, da potência e lidam mal com a heterogeneidade de variâncias. Há alternativas que apresentam os mesmos problemas que a versão paramétrica quando a distribuição não é normal, nem todas são adequados para testar a interação e algumas têm problemas quando as amostras são pequenas. Os métodos que utilizam permutações nem sempre verificam o pressuposto de permutabilidade das observações, ou seja, a probabilidade dos dados observados ser invariante relativamente às permutações aleatórias dos índices.

Na Tabela 1 apresenta-se um resumo das principais características de cada uma destas alternativas encontradas na literatura (e.g. [2, 4, 5, 7, 8]). A avaliação do desempenho destas técnicas, face a diferentes graus de assimetria, presença de *outliers* e heterogeneidade de variâncias, foi realizada com base em distribuições contínuas e considerando delineamentos equilibrados e/ou desequilibrados.

3 Simulação

No estudo de simulação foi considerado um modelo de efeitos fixos e com interação,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

Tabela 1: Principais vantagens e desvantagens das alternativas à ANOVA paramétrica. (+ bom desempenho, – mau desempenho, \pm o desempenho depende de algumas características, n.a. não aplicável)

Método	Fácil	Erro Tipo I	Potência	Dist. não normal	Heterogeneidade	Interação	Software	Permutabilidade	Amostras pequenas
<i>RT</i>	+	–	–	–	–	–	+	n.a.	–
<i>INT</i>	+	–	\pm				+	n.a.	–
<i>ART</i>	–	\pm	\pm	+	–	+	+	n.a.	+
<i>ART+INT</i>	–	\pm	\pm	+	–	–	+	n.a.	\pm
<i>L de PS</i>	+	\pm	\pm	+		+	+	n.a.	+
<i>vdW</i>	–	+	+	+	+	+	+	n.a.	\pm
<i>WTS</i>	–	\pm		+	+		+	n.a.	–
<i>ATS</i>	–	+	–	+	+	+	+	n.a.	+
<i>WTPS</i>	–	+	+	+	+	+	+	\pm	+
<i>CPS</i>	–	–	+				+	–	+
<i>UPS</i>	–	–	+				+	–	+

onde μ é a média global, α_i o efeito do nível i do fator A , $i = 1, \dots, L$, β_j o efeito do nível j do fator B , $j = 1, \dots, C$, γ_{ij} é o efeito da interação do nível i do fator A com o nível j do fator B e ϵ_{ijk} é o erro aleatório, $k = 1, \dots, n$.

Os efeitos principais do fator A foram modelados considerando $\alpha_1 = c$, $\alpha_2 = -c$ e $\alpha_i = 0$ se $i \neq 1, 2$, com $c = 0,25\sigma, 0,5\sigma$ e 1σ onde σ representa o desvio-padrão da população amostrada. Os efeitos principais do fator B foram modelados considerando $\beta_1 = c$, $\beta_2 = -c$

e $\beta_i = 0$ se $i \neq 1, 2$. As interações $A \times B$ foram criadas definindo $\gamma_{11} = \gamma_{22} = c$, $\gamma_{12} = \gamma_{21} = -c$ e $\gamma_{ij} = 0$ nos restantes casos.

As taxas de erro de tipo I e II dos vários testes foram avaliadas considerando dois cenários distintos: (1) um efeito principal e inexistência de interação, ou seja, $c \neq 0$ para o efeito A e $c = 0$ para os outros efeitos; (2) um efeito principal e existência de interação, i.e., $c \neq 0$ para os efeitos A e $A \times B$ e $c = 0$ para o efeito B .

Foram considerados delineamentos equilibrados ($n_{ij} = 3, 5, 10$) com 2 fatores, A e B , com igual e desigual número de níveis ($2 \times 5, 3 \times 3, 3 \times 4$) e dados provenientes de distribuições discretas, com diferentes parâmetros de modo a obter vários graus de dispersão e assimetria: (i) Binomial assimétrica positiva: $B(N; 0, 2)$ com $N = 25, 50, 100$; (ii) Binomial simétrica: $B(N; 0, 5)$ com $N = 10, 20, 40$; (iii) Binomial Negativa: $BN(N; 0, 4)$ com $N = 2, 4, 8$; (iv) Poisson: $P(\lambda)$ com $\lambda = 5, 10, 20$; e (v) Uniforme: $\{0, \dots, N\}$ com $N = 10, 20, 40$.

Para cada cenário distribucional foram realizadas $M = 1000$ replicações tendo-se registado, para cada um dos testes descritos na secção anterior, a distribuição empírica dos valores p , a proporção de réplicas que rejeitaram H_0 quando H_0 verdadeiro (*taxa de erro de tipo I empírica*) e a proporção de réplicas que não rejeitaram H_0 quando H_1 verdadeiro (*taxa de erro de tipo II empírica*), ao nível de significância definido, $\alpha = 1\%, 5\%$ e 10% . Os testes *ART*, *ATS* e *WTS* foram aplicados quer às observações originais (y) quer às respectivas ordens (ry). Para distinguir entre estas duas situações, na apresentação dos resultados utilizaram-se os sufixos y e ry , respetivamente.

Na análise do desempenho dos testes no controlo da probabilidade do erro de tipo I, foi usado o critério liberal de Bradley [1]. Segundo este critério, um teste pode ser considerado robusto se a sua taxa de erro de tipo I empírica estiver no intervalo $[0, 5\alpha; 1, 5\alpha]$. O teste é considerado conservador se a taxa empírica estiver abaixo do limite inferior e é considerado liberal se estiver acima do limite superior.

Foram usados os pacotes ARTTool, rankFD e GFD do programa R Project [9], e funções disponíveis em <http://www.uni-koeln.de/~luepsen/R/> e <http://static.gest.unipd.it/~salmaso/web/>.

4 Resultados

Dado não ser possível mostrar todos os resultados, nas Figuras 1 e 2 ilustra-se o comportamento genérico da distribuição empírica dos *valores p* dos vários testes.

Quando o efeito em estudo não está presente, os testes de permutação *CSP* e *USP* são os que apresentam uma distribuição com maior dispersão (Figuras 1 e 2). Os testes *USP* e *WTS* distinguem-se de todos os restantes por na sua distribuição predominarem valores mais elevados. Os testes *L de PS* e *van der Waerden* destacam-se ora por apresentarem uma frequência maior de *valores p* menores ora pelo comportamento inverso (no teste ao efeito quando a interação está presente).

Quando o efeito em estudo está presente, a distribuição tende a ser assimétrica e a possuir vários valores atípicos (Figuras 1 e 2). A predominância de *valores p* mais elevados é maior nos testes *USP* e *WTS*, seguindo-se os testes *INT*, *ANOVA* e *ART*. Pelo contrário, uma maior frequência de *valores p* mais baixos é registada nos testes *L de PS* e *van der Waerden* (excepto no teste ao efeito isolado quando a interação está presente) e posteriormente nos testes *CSP* e *ATS*.

As taxas de erro de tipo I e II empíricas reportadas nas Tabelas 2 a 4 correspondem à média das taxas de erro produzidas por cada um dos testes em todos os cenários considerados. De acordo com o critério de Bradley [1], os testes *WTS* e *USP* são demasiado liberais. Os testes de *L de PS* e *van der Waerden* apresentam um comportamento instável; são muito conservadores na ausência de interação, mas na presença de interação a sua taxa de erro de tipo I ultrapassa o valor de α . O teste *CSP* é o que apresenta a maior taxa de erro de tipo II empírica. Os restantes testes apresentam taxas de erro empíricas semelhantes, embora o teste *ATS* seja o que menos vezes apresentou taxas de erro de tipo I empíricas superiores ao valor α nominal e o teste *ART* o que mais vezes ultrapassou o valor α .

Na análise que se segue exclui-se-ão os testes *USP*, *WTS*, *L de PS* e *van der Waerden* por violarem o critério de robustez [1].

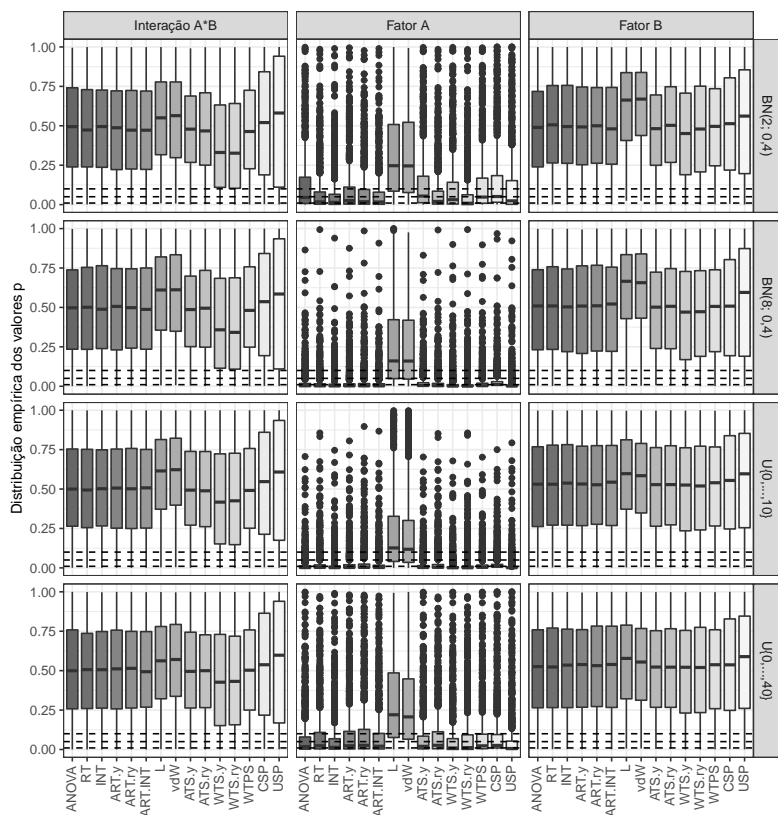


Figura 1: Distribuição empírica dos *valores p*, quando $n = 5$, $c = 0,5\sigma$ para o efeito A , $c = 0$ para os efeitos B e $A \times B$, e $\epsilon_{ijk} \sim BN(N; 0, 4)$ com $N = 2, 8$ e $\epsilon_{ijk} \sim U\{0, \dots, N\}$ com $N = 10, 40$, no delineamento 3×3 . (as linhas horizontais tracejadas representam os níveis de significância de 1%, 5% e 10%)

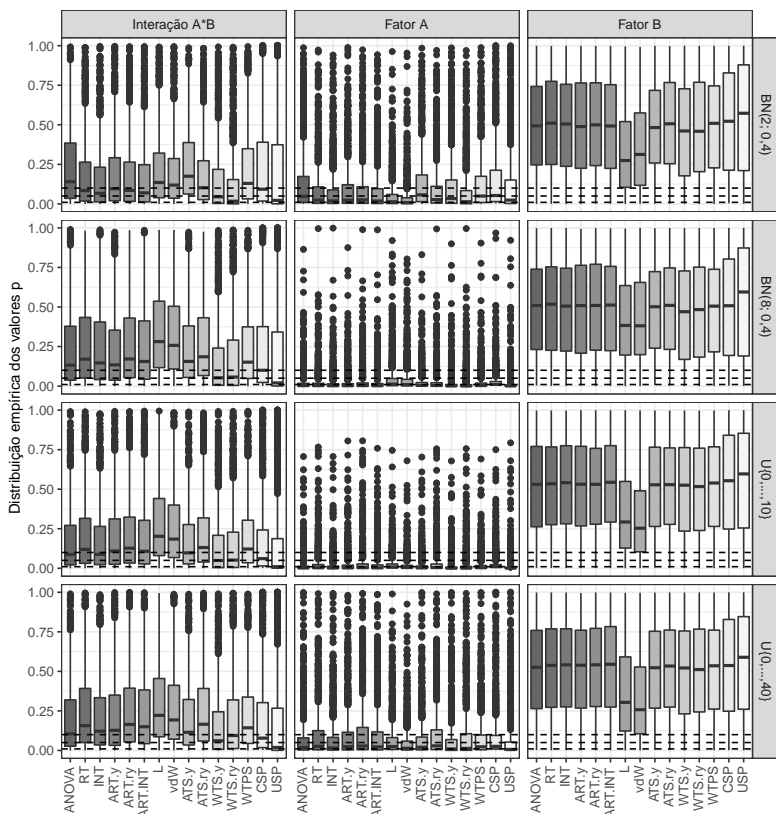


Figura 2: Distribuição empírica dos *valores p*, quando $n = 5$, $c = 0,5\sigma$ para os efeitos A e $A \times B$, $c = 0$ para o efeito B , e $\epsilon_{ijk} \sim BN(N; 0,4)$ com $N = 2, 8$ e $\epsilon_{ijk} \sim U\{0, \dots, N\}$ com $N = 10, 40$, no delineamento 3×3 . (as linhas horizontais tracejadas representam os níveis de significância de 1%, 5% e 10%)

Tabela 2: Média das taxas de erro de tipo I empíricas dos métodos no teste à presença de efeito principal B , para diferentes níveis de significância (α). Os valores de erro de tipo I que mais se afastam do nível de significância definido estão a itálico.

Interação	inexistente			existente		
α	0,01	0,05	0,10	0,01	0,05	0,10
<i>ANOVA</i>	0,010	0,049	0,098	0,010	0,049	0,099
<i>RT</i>	0,011	0,050	0,098	0,010	0,048	0,096
<i>INT</i>	0,010	0,049	0,099	0,009	0,045	0,092
<i>ART.y</i>	0,011	0,051	0,101	0,011	0,051	0,102
<i>ART.ry</i>	0,011	0,051	0,100	0,011	0,050	0,098
<i>ART+INT</i>	0,011	0,050	0,100	0,010	0,046	0,092
<i>L de PS</i>	<i>0,001</i>	<i>0,013</i>	<i>0,036</i>	<i>0,057</i>	<i>0,166</i>	<i>0,260</i>
<i>vdW</i>	<i>0,001</i>	<i>0,014</i>	<i>0,037</i>	<i>0,067</i>	<i>0,181</i>	<i>0,277</i>
<i>ATS.y</i>	0,006	0,037	0,083	0,006	0,037	0,083
<i>ATS.ry</i>	0,007	0,040	0,087	0,006	0,038	0,083
<i>WTS.y</i>	<i>0,036</i>	<i>0,096</i>	<i>0,155</i>	<i>0,037</i>	<i>0,097</i>	<i>0,155</i>
<i>WTS.ry</i>	<i>0,040</i>	<i>0,100</i>	<i>0,157</i>	<i>0,038</i>	<i>0,095</i>	<i>0,151</i>
<i>WTPS</i>	0,010	0,048	0,097	0,010	0,049	0,097
<i>CSP</i>	0,012	0,047	0,110	0,012	0,047	0,109
<i>USP</i>	<i>0,046</i>	<i>0,104</i>	<i>0,155</i>	<i>0,046</i>	<i>0,105</i>	<i>0,156</i>

Análise da distribuição: No teste ao efeito principal, a taxa de erro de tipo II empírica dos vários testes parece ser superior quando a distribuição dos erros é assimétrica e, além disso, a distribuição dos *valores p* apresenta maior dispersão. Quando a distribuição é simétrica, o tipo de achatamento influencia a dispersão dos *valores p* sendo mais elevada na distribuição platicúrtica.

A taxa de erro de tipo I empírica dos vários testes é semelhante qualquer que seja a distribuição considerada.

Análise do efeito do tamanho da amostra: Com amostras de reduzida dimensão ($n = 3$), o teste *ATS* revelou ser conservador

Tabela 3: Média das taxas de erro de tipo II empíricas dos métodos no teste à presença de efeito principal A , para diferentes níveis de significância (α).

Interação	inexistente			existente		
α	0,01	0,05	0,10	0,01	0,05	0,10
<i>ANOVA</i>	0,395	0,276	0,219	0,395	0,277	0,219
<i>RT</i>	0,394	0,278	0,220	0,403	0,283	0,224
<i>INT</i>	0,381	0,264	0,208	0,386	0,268	0,211
<i>ART.y</i>	0,393	0,279	0,223	0,392	0,279	0,222
<i>ART.ry</i>	0,398	0,284	0,227	0,405	0,288	0,230
<i>ART+INT</i>	0,390	0,276	0,219	0,393	0,278	0,221
<i>ATS.y</i>	0,424	0,293	0,230	0,424	0,294	0,230
<i>ATS.ry</i>	0,415	0,288	0,228	0,428	0,295	0,232
<i>WTPS</i>	0,409	0,284	0,224	0,407	0,284	0,224
<i>CSP</i>	0,605	0,485	0,337	0,605	0,485	0,337

quando $\alpha = 1\%$ tal como o teste *CSP* para $\alpha = 1\%$ e 5% .

A taxa de erro de tipo II empírica de todos os testes diminui com o aumento da dimensão da amostra por célula e há um aumento na dispersão dos *valores p*. Além disso, os testes tendem a apresentar uma taxa de erro de tipo II empírica semelhante.

Análise dos efeitos considerados: A média das taxas de erro de tipo I empíricas dos vários testes não se altera com a intensidade do efeito considerado. Quando o efeito não está presente, a percentagem de vezes que os testes *ART*, *INT*, *RT* e *ART+INT* decidem corretamente aumenta com a intensidade do efeito.

A taxa de erro de tipo II empírica de todos os testes diminui com o aumento da intensidade do efeito e os testes tendem a apresentar um comportamento semelhante. Quando se considera um efeito com intensidade $0,25\sigma$, de um modo geral, todos os testes não detectam a existência de interação. Contudo, à medida que se aumenta a intensidade do efeito os *valores p* reduzem, bem como a dispersão

Tabela 4: Média das taxas de erro de tipo I e de tipo II empíricas dos métodos no teste à existência de interação AB , na presença de um efeito principal significativo, para diferentes níveis de significância (α). Os valores de erro de tipo I que mais se afastam do nível de significância definido estão a itálico.

Erro α	tipo I			tipo II		
	0,01	0,05	0,10	0,01	0,05	0,10
<i>ANOVA</i>	0,010	0,049	0,099	0,676	0,535	0,451
<i>RT</i>	0,011	0,050	0,100	0,697	0,556	0,471
<i>INT</i>	0,010	0,049	0,099	0,672	0,533	0,450
<i>ART.y</i>	0,012	0,053	0,104	0,677	0,537	0,453
<i>ART</i>	0,012	0,053	0,102	0,697	0,557	0,473
<i>ART.INT</i>	0,011	0,051	0,101	0,679	0,541	0,458
<i>L de PS</i>	0,002	0,017	0,041	0,849	0,710	0,609
<i>vdW</i>	0,002	0,016	0,041	0,829	0,686	0,583
<i>ATS.y</i>	0,005	0,033	0,077	0,723	0,579	0,487
<i>ATS.ry</i>	0,006	0,036	0,082	0,742	0,597	0,502
<i>WTS.y</i>	0,065	0,141	0,206	0,546	0,424	0,353
<i>WTS.ry</i>	0,079	0,155	0,220	0,528	0,415	0,350
<i>WTPS</i>	0,010	0,049	0,098	0,718	0,573	0,483
<i>CSP</i>	0,016	0,058	0,124	0,715	0,603	0,477
<i>USP</i>	0,093	0,163	0,213	0,458	0,377	0,332

da sua distribuição empírica, verificando-se que quando o efeito é 1σ e $n = 10$ os testes já decidem corretamente.

Análise dos empates: O número de empates depende tanto do número de observações por célula (n) como dos parâmetros das distribuições. Dado que já foi feita a análise do desempenho dos testes à dimensão da amostra, em que quanto maior a dimensão da amostra maior o número de empates presentes nos dados, focar-se-á apenas o efeito da alteração dos parâmetros das distribuições.

Na ausência de interação, o número de empates não altera a taxa de

erro de tipo I empírica.

A taxa de erro de tipo II empírica do teste à presença do efeito principal tende a não ser afectada pela existência, ou não, de interação. O desempenho dos testes *ATS*, *WTPS* e *CPS*, na detecção da presença de interação, não mostrou ser sensível à alteração dos parâmetros das várias distribuições consideradas, a *ANOVA* paramétrica por vezes apresentou ligeiras alterações no seu desempenho não se identificando qualquer padrão, e os restantes testes mostraram ser instáveis.

Análise dos efeitos dos delineamentos: De um modo geral, não se registam diferenças entre delineamentos no comportamento dos testes.

Quando não existe interação, os testes tendem a apresentar uma taxa de erro de tipo II empírica mais baixa no delineamento 2×5 e mais elevada no delineamento 3×3 . Quando a interação está presente, a taxa de erro de tipo II empírica dos testes não é afectada pelo número de níveis dos fatores.

5 Conclusão

Com base nos resultados obtidos neste estudo de simulação, verificou-se que a distribuição empírica dos *valores p* das várias alternativas à *ANOVA*, bem como da *ANOVA* paramétrica, é afectada por vários fatores como sejam, a dimensão da amostra, a intensidade do efeito, a distribuição dos erros, o número de empates e pelo número de níveis dos fatores.

Os testes de permutação *CSP* e *USP* e o teste *WTS* são os que apresentam a maior dispersão na distribuição empírica das taxas de erro de Tipo I. Além disso, estes testes mostraram ser liberais. O comportamento dos testes *L de PS* e de *van der Waerden* não é consistente, tanto são testes conservadores como liberais. Entre os restantes testes, o teste *ATS* é o que apresenta menores taxas de erro

de tipo I empíricas, mas por sua vez é dos que apresenta maior taxa de erro de tipo II empírica.

A taxa de erro de tipo II empírica de todos os testes diminui com o aumento da dimensão da amostra por célula, com o aumento a intensidade do efeito e os testes tendem a apresentar um comportamento semelhante.

Na presença de interação, todos os testes apresentaram um fraco desempenho no teste à interação, apresentando taxas de erro de tipo II elevadas, mas estas tendem a diminuir com o aumento da dimensão da amostra por célula e com a intensidade dos efeitos.

Os testes *INT*, *ANOVA* e *ART* apresentam taxas de erro de tipo I e II empíricas semelhantes. Comportam-se melhor no estudo do efeito principal presente do que no estudo da interação, do que as restantes alternativas não paramétricas.

Com base nos resultados obtidos, não é aconselhável a utilização dos testes *USP*, *WTS*, *L de PS* e *van der Waerden*. Também é desaconselhada a utilização das restantes alternativas para testar a interação na presença de um efeito principal, especialmente quando a intensidade do efeito é pequena e a dimensão da amostra reduzida. De futuro pretendemos estender este estudo considerando delineamentos desequilibrados, variâncias heterogêneas e a existência de células omissas.

Agradecimentos

Este trabalho é financiado por Fundos Nacionais através da FCT - Fundação para a Ciência e a Tecnologia no âmbito do projeto “UID/MAT/04674/2019 (CIMA)”.

Referências

- [1] Bradley, J. V. (1978). Robustness? *British Journal of Mathematics and Statistical Psychology*, 31, 144–151.

- [2] Hahn, S., Konietzschke, F., Salmaso, L. (2014). A Comparison of efficient permutation tests for unbalanced ANOVA in two by two designs and their behavior under heteroscedasticity. In Melas V., Mignani S., Monari P., Salmaso L. (eds.): *Topics in Statistical Simulation. Springer Proceedings in Mathematics & Statistics*, 114, 257–269.
- [3] Kaptein, M., Nass, C., Markopoulos, P. (2010). Powerful and consistent analysis of Likert-type rating scales. *Proceedings of CHI 2010*, 2391–2394
- [4] Luepsen, H. (2017). The aligned rank transform and discrete variables - a warning. *Communications in Statistics - Simulation and Computation*, 46, 6923–6936.
- [5] Mansouri, H., Chang, G.-H. (1995). A comparative study of some rank tests for interaction. *Computational Statistics & Data Analysis*, 19, 85–96.
- [6] Mansouri, H., Paige, R., Surles, J. G. (2004). Aligned rank transform techniques for analysis of variance and multiple comparisons. *Communications in Statistics - Theory and Methods*, 33, 2217–2232.
- [7] Pauly, M., Brunner, E., Konietzschke, F. (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society, Series B*, 77, 461–473.
- [8] Toothaker, L. E., Newman, D. (1994). Nonparametric competitors to the two-way ANOVA. *Journal of Educational Statistics*, 19, 237–273.
- [9] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Uma nova abordagem na avaliação da interacção genótipo×ambiente em espécies lenhosas de propagação vegetativa: o caso de clones de videira

Elsa Gonçalves

Secção de Matemática/DCEB e LEAF, Instituto Superior de Agronomia, Universidade de Lisboa, *elsagoncalves@isa.ulisboa.pt*

Antero Martins

LEAF, Instituto Superior de Agronomia, Universidade de Lisboa, *anteromart@isa.ulisboa.pt*

Palavras-chave: Dependência nos erros aleatórios; Medidas repetidas; Modelo autoregressivo de primeira ordem; Modelo de simetria composta; Modelos mistos.

Resumo: Neste trabalho estuda-se a interacção genótipo×ambiente ($G \times E$) numa espécie perene. Neste contexto, fazem-se avaliações num dado local na mesma unidade experimental ao longo de anos consecutivos. Ou seja, os erros aleatórios associados a duas quaisquer observações na mesma unidade experimental não são independentes. Ajustam-se, aos dados de rendimento, modelos mistos com matrizes de covariâncias do vector dos erros aleatórios distintas, como a matriz de simetria composta (CS) e autorregressiva de primeira ordem (AR1), de modo a caracterizar esse fenómeno. Verifica-se que os modelos com matrizes CS e AR1 revelam um melhor ajustamento face ao modelo que admite erros aleatórios independentes, garantindo assim o estudo mais preciso da interacção $G \times E$.

1 Introdução

A avaliação da interacção genótipo×ambiente ($G \times E$) é um objectivo incontornável de qualquer programa de melhoramento de plan-

tas para obtenção de variedades geneticamente homogéneas (todos os indivíduos geneticamente iguais). Para compreender o comportamento de um genótipo em ambientes distintos é essencial que a avaliação das características alvo seja feita no maior número de ambientes possível. Em melhoramento de espécies anuais e perenes herbáceas a avaliação da interação é uma prática corrente [6, 9, 2], dada a reduzida área e duração dos ensaios. Sobre este tema, as principais técnicas de estudo deste fenómeno são abordadas em Gonçalves e Martins [4]. Relativamente a espécies lenhosas de propagação vegetativa, em Portugal tem sido feito um esforço coerente para o estudo da interação $G \times E$ em clones de videira. As várias técnicas de interpretação do fenómeno estudadas têm sido aplicadas a dados de rendimento (kg de uva/planta) destacando-se: representação gráfica da ordenação dos clones quanto a diferentes características nos diversos ambientes, cálculo do coeficiente de variação do rendimento de um genótipo nos distintos ambientes, análise de regressão dos valores do rendimento de um genótipo sobre os índices ambientais [5], ajustamento de modelos mistos multivariados e construção de *biplots* [4, 3]. No entanto, tratando-se de uma planta perene (ciclo de vida plurianual), outras abordagens poderão ser ainda exploradas.

Neste tipo de espécie, a instalação de ensaios multi-locais é um processo demorado e de custos elevados, pelo que, na prática, o número de ensaios instalados para o efeito raramente é o mais desejável. De facto, o tempo de vida útil de uma vinha comercial é da ordem dos 30 anos e, por ser uma espécie arbustiva, a área ocupada pelo ensaio tem significado em termos de encargos de gestão cultural. Frequentemente instalam-se ensaios em 2-4 locais e procura-se compensar esse reduzido número com avaliações em mais anos em cada local. A forma mais simples de tratar o problema passa pelo ajustamento de modelos mistos que admitam que os erros aleatórios associados a observações feitas em anos diferentes na mesma unidade experimental (ou seja, sobre as mesmas plantas) são variáveis aleatórias independentes e identicamente distribuídas. No entanto, o que se tem na prática no mesmo local são avaliações sucessivas nos mesmos

indivíduos de cada genótipo ao longo de vários anos. Ou seja, os erros aleatórios associados a duas quaisquer observações na mesma unidade experimental não são independentes. Contudo, para algumas características poderá pensar-se que tal terá pouco significado prático. Por exemplo, uma das características mais importantes alvo do estudo da interacção $G \times E$, o rendimento, é uma característica que é avaliada de ano a ano. Os factores que a influenciam são de tal ordem numerosos que, comparativamente a outras causas de variação, a correlação que existe entre observações feitas na mesma unidade experimental tende a ser negligenciada. No entanto, é necessário compreender se tal se verifica e em que medida afecta a avaliação da interacção $G \times E$. Este trabalho tem precisamente como objectivo desenvolver uma abordagem que responda a este problema.

2 Modelos

Neste trabalho a metodologia proposta, também aplicável a outras espécies lenhosas de propagação vegetativa, representa uma nova abordagem para o estudo da interacção $G \times E$ em clones de videira. Trata-se de uma análise que permite obter os valores médios dos genótipos para uma determinada característica a nível global dos ambientes, juntamente com os desvios da interacção $G \times E$ para cada ambiente e que toma em conta que os erros aleatórios associados a observações feitas em anos diferentes na mesma unidade experimental não são independentes.

Como já referido, nesta espécie em um mesmo local uma determinada característica é avaliada durante vários anos. A noção de ambiente está, portanto, hierarquizada: por um lado, o local distinto (que abrange as condições edafo-climáticas, porta-enxerto, etc.) e, dentro do local, os diferentes anos (não necessariamente os mesmos anos em cada local). Poderia ser construído um modelo que incluisse os efeitos do local e do ano subordinado ao local. No entanto, para simplificação e com o objectivo de centrar o estudo no principal efeito

global do ambiente, nesta abordagem será considerado no modelo o efeito do ambiente visto como a combinação local/ano.

Admitamos dados de rendimento (kg/planta) provenientes de ensaios com delineamento experimental em blocos completos casualizados. Matricialmente, o modelo linear misto aplicável a este tipo de estudo pode ser genericamente descrito como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (1)$$

em que:

\mathbf{Y} é o vector $n \times 1$ das observações (valores fenotípicos, ou observados, do rendimento), ordenado por local, ambiente (combinação local/ano) e unidade experimental dentro de cada ambiente;

\mathbf{X} é a matriz de delineamento $n \times p$ dos efeitos fixos (matriz cujas colunas são variáveis indicatrizes que identificam as observações de cada nível de cada factor de efeitos fixos);

$\boldsymbol{\beta}$ é o vector $p \times 1$ de efeitos fixos (média populacional, efeitos dos ambientes);

\mathbf{Z} é a matriz de delineamento $n \times q$ dos efeitos aleatórios (matriz cujas colunas são variáveis indicatrizes que identificam as observações de cada nível de cada factor de efeitos aleatórios);

\mathbf{u} é o vector $q \times 1$ de efeitos aleatórios que contém os efeitos genotípicos, os efeitos dos blocos por ambiente e os efeitos da interacção genótipo \times ambiente ($q = \sum_{i=1}^r q_i$, sendo q_i o número de níveis do factor de efeitos aleatórios i e r o número de factores de efeitos aleatórios em estudo);

\mathbf{e} é o vector $n \times 1$ de erros aleatórios.

Os vectores \mathbf{u} e \mathbf{e} admitem-se independentes, com distribuição normal multivariada de vector de valores médios nulo e matrizes de covariâncias \mathbf{G} e \mathbf{R} , respectivamente, isto é,

$$\text{Cov}[\mathbf{u}, \mathbf{e}] = \mathbf{0}, \quad \mathbf{u} \sim \mathcal{N}_q(\mathbf{0}, \mathbf{G}), \quad \mathbf{e} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{R}).$$

A distribuição de \mathbf{Y} admite-se assim normal multivariada, com vector de valores médios $\mathbf{X}\boldsymbol{\beta}$ e matriz de covariâncias $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$,

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}).$$

No contexto biológico em análise, estudam-se genótipos sem relações de parentesco e o delineamento experimental é específico de cada ensaio. Assim, relativamente ao vector \mathbf{u} , sendo \mathbf{u}_i o vector dos efeitos aleatórios do factor i , admite-se:

$\text{var}[\mathbf{u}_i] = \mathbf{G}_i = \sigma_{u_i}^2 \mathbf{I}_{q_i}$, para $i = 1, \dots, r$, e $\text{Cov}[\mathbf{u}_i, \mathbf{u}_{i'}] = \mathbf{0}$, para $\forall i \neq i'$.

Consequentemente, a matriz de covariâncias do vector \mathbf{u} é definida como $\mathbf{G} = \oplus_{i=1}^r \mathbf{G}_i$, em que \oplus representa a soma directa de matrizes. Especificando de acordo com efeitos aleatórios descritos no modelo (1), \mathbf{G} é a soma directa das matrizes $\mathbf{G}_g = \sigma_g^2 \mathbf{I}_{q_1}$, $\mathbf{G}_{b(A)} = \sigma_{b(A)}^2 \mathbf{I}_{q_2}$, $\mathbf{G}_{ge} = \sigma_{ge}^2 \mathbf{I}_{q_3}$, sendo σ_g^2 a variância genotípica, $\sigma_{b(A)}^2$ a variância dos blocos subordinados ao ambiente e σ_{ge}^2 a variância da interacção $\mathbf{G} \times \mathbf{E}$.

Relativamente ao vector \mathbf{e} , a forma mais simples de abordar o problema é admitir que os elementos deste vector são variáveis aleatórias independentes e identicamente distribuídas, isto é, $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$. Admite-se, portanto, homogeneidade de variâncias e que os erros aleatórios associados a observações feitas em anos diferentes na mesma unidade experimental (em ambientes diferentes no mesmo local) são variáveis aleatórias independentes e identicamente distribuídas (adiante designado modelo *IND*). No entanto, o que se tem na prática no mesmo local são avaliações sucessivas na mesma unidade experimental ao longo de vários anos.

Seja, então, o vector \mathbf{e} definido como

$$\mathbf{e} = \begin{bmatrix} \mathbf{e}_{1(n_1 \times 1)}^T & \mathbf{e}_{2(n_2 \times 1)}^T & \cdots & \mathbf{e}_{l(n_l \times 1)}^T \end{bmatrix}^T,$$

representando cada subvector, \mathbf{e}_j , com $j = 1, \dots, l$, o vector dos erros aleatórios no local j , e n_j o número de observações no respectivo local. Seja $\text{Var}[\mathbf{e}_j] = \mathbf{R}_j$, representando, portanto, \mathbf{R}_j a respectiva matriz de covariâncias. Admite-se $\text{Cov}[\mathbf{e}_j, \mathbf{e}_{j'}] = \mathbf{0}$, para $\forall j \neq j'$,

consequentemente, a matriz de covariâncias do vector \mathbf{e} é definida como

$$\mathbf{R} = \oplus_{j=1}^r \mathbf{R}_j.$$

Trata-se agora de definir a estrutura da matriz de covariâncias do vector dos erros aleatórios no local j (matriz \mathbf{R}_j). Admite-se que os erros aleatórios associados a observações de unidades experimentais diferentes são independentes e que os associados a observações na mesma unidade experimental ao longo dos anos não são independentes. No local j com p unidades experimentais, a matriz \mathbf{R}_j é definida como

$$\mathbf{R}_j = \mathbf{I}_p \otimes \Sigma_{e_j},$$

sendo \mathbf{I}_p a matriz identidade de ordem p e \otimes o produto de *Kronecker* de matrizes.

A estrutura mais complexa para Σ_{e_j} seria uma matriz não estruturada, que admitiria variâncias distintas para cada ano no local j e diferentes covariâncias para todos os pares de anos avaliados nesse local. Por local, para a anos de avaliação, estimar-se-iam $a + a(a-1)/2$ parâmetros, o que resultaria num modelo excessivamente parametrizado e um exagero em termos do que realmente se quer ter em conta [12]. Não será, portanto, a opção a seguir. As estruturas estudadas serão as que fazem sentido neste contexto biológico: uma estrutura que traduza uma contribuição comum a todas as observações feitas na mesma unidade experimental, ou que considere que a correlação entre observações da mesma unidade experimental diminui à medida que a distância de separação entre os anos aumenta. Neste sentido, para a matriz Σ_{e_j} foram definidos os dois tipos de estrutura seguidamente descritos.

(1) Uma matriz de simetria composta (adiante designado modelo CS, do Inglês *Compound Symmetry*), que tem como elementos diagonais $\sigma_{e_j}^2$ (variância dos erros aleatórios do local j) e elementos não diagonais $\sigma_{e_j}^2 \rho$ (sendo ρ a correlação entre pares de observações na mesma unidade experimental ao longo dos anos). Isto é, admite-se homogeneidade de variâncias no local j e que todos os pares de observações na mesma unidade experimental têm a mesma correlação.

(2) Uma matriz autorregressiva de primeira ordem (adiante designado modelo AR1), que tem como elementos diagonais $\sigma_{e_j}^2$ e elementos não diagonais $\sigma_{e_j}^2 \rho^{|k-k'|}$, sendo $|k - k'|$ o intervalo de separação entre os anos k e k' . Isto é, admite-se homogeneidade de variâncias no local j e que a correlação entre observações da mesma unidade experimental diminui à medida que a distância entre anos aumenta. Faz sentido quando se avaliam anos consecutivos.

O método de estimação dos parâmetros incluídos nas matrizes **G** e **R** foi o método de máxima verosimilhança restrita, REML [8], actualmente o mais recomendado e utilizado no contexto dos modelos lineares mistos para estimar parâmetros covariância em grandes conjuntos de dados com estrutura complexa [7]. A comparação e selecção de modelos com estruturas de covariância distintas (modelos IND, CS e AR1) foi baseada no critério de informação de Akaike (AIC) [10], definido como [12]:

$$AIC_i = -2lr_i + 2npar_i,$$

em que lr_i é a log-verosimilhança restrita do modelo i e $npar_i$ o respectivo número de parâmetros covariância do modelo. Modelos encaixados foram também formalmente comparados com base em testes de razão de verosimilhanças.

Por último, é útil reforçar que esta metodologia é aplicada na última fase da metodologia de selecção da videira [5], quando todos os genótipos (clones) presentes já foram seleccionados quanto a várias características de interesse, mas ainda não quanto à sua estabilidade ambiental. Por isso, o objectivo principal nesta fase é avaliar a variabilidade da interacção $G \times E$ e, posteriormente, seleccionar com base na menor sensibilidade a essa interacção. Neste sentido, o interesse deste tipo de estudo está essencialmente nesta componente de variância (σ_{ge}^2). A inferência relativa a este parâmetro foi baseada num teste de razão de verosimilhanças restritas [11] ($H_0 : \sigma_{ge}^2 = 0$ vs $H_0 : \sigma_{ge}^2 > 0$) e foi efectuada para todos os modelos em estudo,

de modo a compreender em que medida a estrutura de covariância dos erros aleatórios afecta a avaliação da interacção $G \times E$.

3 Uma aplicação

A aplicação utiliza dados de rendimento (kg de uva/planta) de vários ensaios de selecção de variedades antigas de videira, instalados em 2 e 3 locais por variedade, segundo um delineamento experimental em blocos completos casualizados (RCB, do Inglês *Randomized Complete Block*), com 3 a 9 repetições, e avaliados durante vários anos (Tabela 1). O delineamento experimental RCB é um dos mais simples para controlar variabilidade indesejável presente num ensaio (tipo de solo, exposição, declive, etc.). Neste sentido, nos modelos de análise de dados de ensaios com este tipo de delineamento, os efeitos dos blocos por ambiente são sempre incluídos, de modo a respeitar o processo de casualização associado a este tipo de delineamento.

Procedeu-se de seguida ao ajustamento dos modelos lineares mistos propostos na secção anterior, considerando-se: o factor ambiente como um factor de efeitos fixos; os efeitos genotípicos, os efeitos dos blocos por ambiente e os efeitos da interacção genótipo \times ambiente como aleatórios; as três estruturas de covariância (IND, CS and AR1). Para tal, recorreu-se ao *Software R, package ASReml – R* [1] (método de estimação REML, usando o algoritmo de informação média).

Os resultados obtidos com o ajustamento dos vários modelos aos dados de rendimento das 7 variedades antigas constam da Tabela 2. Em todos os casos estudados verificou-se que os modelos com estruturas de covariância CS e AR1 revelaram melhor ajustamento face ao modelo que admitiu erros aleatórios independentes entre observações da mesma unidade experimental (IND), tendo este último modelo revelado sempre maior valor de *AIC*. Também se chega a esta conclusão comparando formalmente os modelos IND e CS e os modelos IND e AR1 através de um teste de razão de verosimilhanças restritas. Em qualquer dos casos, pelos valores da log-verosimilhança

Alvarinho (AI), 40 clones avaliados em 20 ambientes: <i>l1</i> , Monção (RCB, 3 blocos)/1990 a 1992; <i>l2</i> , Monção-Pias (RCB, 9 blocos)/1995 a 2000; <i>l3</i> , Monção-Ceivães (RCB, 9 blocos)/1994 a 2004.
Antão Vaz (AN), 40 clones avaliados em 14 ambientes: <i>l1</i> , Évora (RCB, 5 blocos)/1988;1989; 1990; <i>l2</i> , Palmela (RCB, 8 blocos)/1993 a 1998; <i>l3</i> , Vidigueira (RCB, 8 blocos)/1998 a 2002.
Aragonez (RZ), 40 clones avaliados em 13 ambientes: <i>l1</i> , Estremoz (RCB, 8 blocos)/1992 a 1999; <i>l2</i> , Tabuaço (RCB, 8 blocos)/1994 a 1998.
Fernão Pires (FP), 40 clones avaliados em 11 ambientes: <i>l1</i> , Alpiarça (RCB, 8 blocos)/1993 a 1997; <i>l2</i> , Anadia (RCB, 8 blocos)/1999,2000; <i>l3</i> , Caldas da Rainha (RCB, 8 blocos)/1995 a 1998.
Negra Mole (NM), 40 clones avaliados em 7 ambientes: <i>l1</i> , Lagoa (RCB, 5 blocos)/1989, 1990; <i>l2</i> , Loulé (RCB, 8 blocos)/1994 a 1998.
Rabo de Ovelha (OV), 33 clones avaliados em 8 ambientes: <i>l1</i> , Redondo (RCB, 8 blocos)/1996 a 2000; <i>l2</i> , Reguengos de Monsaraz (RCB, 5 blocos)/1990 a 1992.
Síria (CR), 40 clones avaliados em 10 ambientes: <i>l1</i> , Estremoz (RCB, 8 blocos)/1992 a 1999; <i>l2</i> , Pinhel (RCB, 5 blocos)/1988 e 1989.

Tabela 1: Descrição dos dados usados na aplicação: variedade antiga, número de clones (genótipos) e número de ambientes por variedade, com indicação do local, delineamento experimental e anos de avaliação em cada local.

Variedade antiga	Modelo	lr	$npar$	AIC
Alvarinho (AI)	<i>IND</i>	-5509.5	4	11027.0
	<i>CS</i>	-5202.5	9	10423.0
	<i>AR1</i>	-5224.9	9	10467.8
Antao Vaz (AN)	<i>IND</i>	-3506.1	4	7020.2
	<i>CS</i>	-3155.9	9	6329.9
	<i>AR1</i>	-3167.7	9	6353.3
Aragonez (RZ)	<i>IND</i>	-1820.8	4	3649.6
	<i>CS</i>	-1573.4	7	3160.9
	<i>AR1</i>	-1711.0	7	3435.9
Fernaõ Pires (FP)	<i>IND</i>	-1548.6	4	3105.3
	<i>CS</i>	-1480.9	9	2979.9
	<i>AR1</i>	-1493.3	9	3004.5
Negra Mole (NM)	<i>IND</i>	-738.6	4	1485.2
	<i>CS</i>	-535.1	7	1084.3
	<i>AR1</i>	-531.2	7	1076.4
Rabo de Ovelha (OV)	<i>IND</i>	-1173.6	4	2355.2
	<i>CS</i>	-1156.5	7	2326.9
	<i>AR1</i>	-1164.4	7	2342.8
Sória (CR)	<i>IND</i>	-1303.9	4	2615.9
	<i>CS</i>	-1250.1	7	2514.2
	<i>AR1</i>	-1285.7	7	2585.5

Tabela 2: Log-verossimilhança restrita (lr), número de parâmetros covariância ($npar$) e critério de informação de Akaike (AIC) obtidos com o ajustamento dos modelos com matrizes de covariâncias do vector dos erros aleatórios diagonal (IND), de simetria composta (CS) e autorregressiva de primeira ordem (AR1).

restrita constantes da Tabela 2, facilmente se percebe que o valor calculado da estatística do teste de razão de verosimilhanças restritas conduzirá à rejeição do modelo IND para qualquer nível de significância usual. Entre os modelos que admitem dependência entre observações realizadas na mesma unidade experimental, o modelo CS revelou sempre um melhor ajustamento, com excepção de um único caso de estudo, a casta Negra Mole (Tabela 2). Os resultados obtidos com o modelo AR1, que não evidenciam uma vantagem geral deste modelo face ao CS, podem dever-se ao facto do número de anos avaliados em cada local ser baixo ou moderado.

As estimativas dos parâmetros covariância obtidas com o ajustamento dos vários modelos encontram-se na Tabela 3. Torna-se clara a heterogeneidade de variâncias do erro entre locais. Relativamente às correlações entre observações realizadas na mesma unidade experimental, estas revelaram-se fracas a moderadas, variando entre castas e, relativamente a algumas castas, também entre locais (diferenças mais marcantes nas castas Alvarinho, Antão Vaz e Negra Mole). A variação dessa correlação entre locais da mesma casta pode dever-se às condições edafo-climáticas e culturais específicas de cada local. Relativamente às estimativas da variância genotípica e da variância dos blocos subordinados ao ambiente, observa-se que com o ajustamento dos modelos CS e AR1 os valores dessas estimativas são, em geral, menores. Pelo contrário, com o ajustamento dos modelos CS e AR1 a estimativa da variância da interacção $G \times E$ tende, em geral, a ser maior.

Quando se testa a componente de variância da interacção $G \times E$ através de um teste de razão de verosimilhanças restritas (Tabela 4), conclui-se que essa variabilidade é significativa, para qualquer nível de significância usual, em todos os casos estudados, excepto no caso Negra Mole com o modelo IND. Verifica-se também que a diferença entre log-verosimilhanças restritas dos modelos com e sem efeitos da interacção é mais marcante nos modelos CS e AR1, resultando num maior valor calculado da estatística do teste de razão de verosimilhanças restritas (REMLRT). No caso da casta Negra Mole a variabilidade da interacção $G \times E$ apenas se revelou signifi-

Estimativas	AI	AN	RZ	FP	NM	OV	CR
Modelo IND							
$\hat{\sigma}_g^2 (SE)$	0.936 (0.246)	0.041 (0.015)	0.054 (0.015)	0.091 (0.027)	0.104 (0.027)	0.381 (0.115)	0.056 (0.017)
$\hat{\sigma}_{b(A)}^2 (SE)$	0.233 (0.046)	0.506 (0.083)	0.177 (0.029)	0.205 (0.035)	0.092 (0.024)	0.737 (0.165)	0.145 (0.029)
$\hat{\sigma}_{ge}^2 (SE)$	0.883 (0.090)	0.072 (0.021)	0.046 (0.010)	0.130 (0.017)	0.010 (0.010)	0.321 (0.050)	0.084 (0.014)
$\hat{\sigma}_e^2 (SE)$	3.684 (0.087)	1.830 (0.044)	0.775 (0.018)	0.758 (0.020)	0.689 (0.024)	1.184 (0.046)	0.757 (0.021)
Modelo CS							
$\hat{\sigma}_g^2 (SE)$	0.797 (0.215)	0.013 (0.012)	0.037 (0.014)	0.088 (0.026)	0.067 (0.024)	0.367 (0.114)	0.040 (0.016)
$\hat{\sigma}_{b(A)}^2 (SE)$	0.161 (0.037)	0.468 (0.078)	0.164 (0.027)	0.202 (0.035)	0.064 (0.018)	0.729 (0.164)	0.138 (0.028)
$\hat{\sigma}_{ge}^2 (SE)$	0.808 (0.077)	0.100 (0.018)	0.065 (0.009)	0.123 (0.016)	0.041 (0.009)	0.312 (0.048)	0.093 (0.014)
$\hat{\sigma}_{e11}^2 (SE)$	1.251 (0.139)	1.842 (0.112)	0.819 (0.030)	0.985 (0.041)	0.868 (0.068)	1.243 (0.056)	0.741 (0.024)
$\hat{\rho}_{11} (SE)$	0.423 (0.072)	0.089 (0.046)	0.273 (0.024)	0.070 (0.024)	0.319 (0.067)	0.100 (0.028)	0.142 (0.020)
$\hat{\sigma}_{e12}^2 (SE)$	5.862 (0.223)	2.496 (0.108)	0.702 (0.030)	0.688 (0.044)	0.654 (0.034)	1.012 (0.089)	0.875 (0.067)
$\hat{\rho}_{12} (SE)$	0.087 (0.023)	0.357 (0.027)	0.284 (0.029)	0.059 (0.064)	0.457 (0.030)	0.211 (0.066)	0.110 (0.075)
$\hat{\sigma}_{e13}^2 (SE)$	2.571 (0.092)	1.026 (0.042)		0.572 (0.024)			
$\hat{\rho}_{13} (SE)$	0.189 (0.023)	0.190 (0.027)		0.130 (0.027)			
Modelo AR1							
$\hat{\sigma}_g^2 (SE)$	0.842 (0.224)	0.019 (0.012)	0.048 (0.014)	0.089 (0.026)	0.080 (0.025)	0.394 (0.118)	0.054 (0.017)
$\hat{\sigma}_{b(A)}^2 (SE)$	0.168 (0.038)	0.464 (0.077)	0.170 (0.028)	0.203 (0.035)	0.070 (0.019)	0.734 (0.165)	0.144 (0.029)
$\hat{\sigma}_{ge}^2 (SE)$	0.805 (0.078)	0.091 (0.017)	0.058 (0.009)	0.122 (0.016)	0.037 (0.009)	0.322 (0.050)	0.090 (0.014)
$\hat{\sigma}_{e11}^2 (SE)$	1.247 (0.130)	1.850 (0.113)	0.813 (0.025)	0.984 (0.040)	0.873 (0.068)	1.234 (0.054)	0.735 (0.022)
$\hat{\rho}_{11} (SE)$	0.394 (0.066)	0.106 (0.055)	0.232 (0.019)	0.075 (0.030)	0.322 (0.067)	-0.109 (0.033)	0.120 (0.021)
$\hat{\sigma}_{e12}^2 (SE)$	5.840 (0.219)	2.479 (0.096)	0.697 (0.027)	0.688 (0.044)	0.606 (0.027)	1.009 (0.086)	0.882 (0.067)
$\hat{\rho}_{12} (SE)$	0.056 (0.028)	0.438 (0.020)	0.271 (0.024)	0.059 (0.064)	0.485 (0.022)	0.159 (0.065)	0.116 (0.075)
$\hat{\sigma}_{e13}^2 (SE)$	2.607 (0.089)	1.029 (0.041)		0.570 (0.023)			
$\hat{\rho}_{13} (SE)$	0.305 (0.021)	0.258 (0.029)		0.110 (0.030)			

Tabela 3: Estimativas dos parâmetros covariância (e respectivos erros padrão, SE) resultantes do ajustamento dos modelos com matrizes de covariâncias do vector dos erros aleatórios diagonal (IND), de simetria composta (CS) e autorregressiva de primeira ordem (AR1): $\hat{\sigma}_g^2$ - estimativa da variância genotípica; $\hat{\sigma}_{b(A)}^2$ - estimativa da variância dos blocos subordinados ao ambiente; $\hat{\sigma}_{ge}^2$ - estimativa da variância da interacção $G \times E$, $\hat{\sigma}_{e_{li}}^2$ - estimativas da variância dos erros aleatórios, $\hat{\rho}_{li}$ - estimativas das correlações entre observações feitas na mesma unidade experimental (no modelo AR1, correlações entre observações feitas na mesma unidade experimental em dois anos consecutivos).

Variedade antiga	Modelo	REMLRT	Valor-p
Alvarinho (AI)	<i>IND</i>	238.5	<0.0001
	<i>CS</i>	253.0	<0.0001
	<i>AR1</i>	257.2	<0.0001
Antao Vaz (AN)	<i>IND</i>	15.9	<0.0001
	<i>CS</i>	57.1	<0.0001
	<i>AR1</i>	51.4	<0.0001
Aragonez (RZ)	<i>R = IND</i>	35.0	<0.0001
	<i>CS</i>	97.7	<0.0001
	<i>AR1</i>	71.0	<0.0001
Fernaio Pires (FP)	<i>IND</i>	153.5	<0.0001
	<i>CS</i>	157.4	<0.0001
	<i>AR1</i>	147.3	<0.0001
Negra Mole (NM)	<i>IND</i>	1.0	0.3142
	<i>CS</i>	37.7	<0.0001
	<i>AR1</i>	36.1	<0.0001
Rabo de Ovelha (OV)	<i>IND</i>	125.2	<0.0001
	<i>CS</i>	126.5	<0.0001
	<i>RAR1</i>	123.0	<0.0001
Síria (CR)	<i>IND</i>	64.6	<0.0001
	<i>CS</i>	95.9	<0.0001
	<i>AR1</i>	81.2	<0.0001

Tabela 4: Teste de razão de verossimilhanças restritas à componente de variância da interação genótipo \times ambiente ($H_0 : \sigma_{ge}^2 = 0$ vs $H_0 : \sigma_{ge}^2 > 0$) para cada um dos modelos com matrizes de covariâncias do vector dos erros aleatórios diagonal (*IND*), de simetria composta (*CS*) e autorregressiva de primeira ordem (*AR1*) e respectivo valor-p. REMLRT, valor calculado da estatística do teste de razão de verossimilhanças restritas.

cativa com os ajustamento destes últimos modelos. Este resultado reforça a vantagem da utilização da abordagem proposta no estudo da interacção $G \times E$. Finalmente, deve-se acrescentar que, inerente a este resultado, está a consequente vantagem na posterior utilização dos melhores preditores empíricos lineares não enviesados dos efeitos da interacção $G \times E$, obtidos com o ajustamento dos modelos CS e AR1, para fins de selecção de clones com menor sensibilidade a essa interacção.

4 Conclusões

Analisando dados de rendimento, os modelos que admitiram que observações na mesma unidade experimental não são independentes (modelos CS e AR1), revelaram melhor ajustamento face ao modelo que admitiu erros independentes (modelo IND), afectando, consequentemente, a avaliação da interacção $G \times E$. De entre as estruturas de covariância estudadas, o modelo CS revelou-se quase sempre como o mais adequado. Isto revela que em ensaios com videira a estrutura que mais se adequa é a que traduz a existência de um efeito comum, ainda que pequeno a moderado, a todas as observações feitas na mesma unidade experimental (o solo que partilha, a estrutura radicular, etc.).

Resumindo, este trabalho dá resposta ao modelo que deverá ser usado no estudo da interacção $G \times E$ na fase final da metodologia de selecção da videira. Este conhecimento é fundamental, pois é o primeiro passo para alcançar o objectivo final: selecção dos clones com menor sensibilidade à interacção $G \times E$. Essa selecção deverá basear-se nos melhores preditores empíricos lineares não enviesados dos efeitos da interacção $G \times E$ obtidos com o ajustamento desse modelo (quanto mais próximos de zero, menor a sensibilidade à interacção).

Agradecimentos

Aos colegas da “Rede Nacional de Selecção da Videira” e da Associação Portuguesa para a Diversidade da Videira (PORVID) pela sua contribuição em todo o processo de selecção das castas. À Fundação para a Ciência e Tecnologia (UID/AGR/04129/2013) pelo apoio financeiro.

Referências

- [1] Butler, D., Cullis, B.R., Gilmour, A.R., Gogel, B.J. (2007). ASReml-R reference manual. ASReml-R estimates variance components under a general linear mixed model by residual maximum likelihood (REML). NSW Department of Primary Industries, Queensland Government. Queensland.
- [2] De Faveri, J., Verbyla, A.P., Pitchford, W.S., Venkatanagappa, S., Cullis, B.R.. Statistical methods for analysis of multi-harvest data from perennial pasture variety selection trials. *Crop and Pasture Science*, 66, 947–962, 2015.
- [3] Gonçalves, E., Carrasquinho, I., Almeida, R., Pedroso, V., Martins, A.. Genetic correlations in grapevine and their effects on selection. *Australian Journal of Grape and Wine Research*, 22, 52–63, 2016.
- [4] Gonçalves, E., Martins, A.. Metodologias estatísticas para estudo da interacção genótipo×ambiente em clones de videira. *Atas XXI Congresso da Sociedade Portuguesa de Estatística*, 89–103, 2014.
- [5] Martins, A., Gonçalves, E.. Grapevine breeding programmes in Portugal. In *Grapevine Breeding Programs for the Wine Industry*. A. G. Reynolds ed., Woodhead Publishing, Elsevier, UK, 159–182, 2015.
- [6] Mathews, K.L., Chapman, S.C., Trethowan, R., Pfeiffer, W., Ginkel, M., Crossa, J., Payne, T., DeLacy, I., Fox, P.N., Cooper, M.. Global adaptation patterns of Australian and CIMMYT spring bread wheat. *Theor Appl Genet*, 115, 819–855, 2007.
- [7] McCulloch, C.E., Searle, S.R., Neuhaus, J.M. (2008). Generalized, linear and mixed models. John Wiley & Sons, New York.

- [8] Patterson, H.D., Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545–554.
- [9] Piepho, H.P., Eckl, T.. Analysis of series of variety trials with perennial crops. *Grass and Forage Science*, 69, 431–440, 2014.
- [10] Sakamoto, Y., Ishiguro, M., Kitagawa, G. (1986). Akaike information criterion statistics. Dordrecht: D. Reidel.
- [11] Stram, D.O., J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* 50, 1171–1177.
- [12] Stroup, W.W. (2013). Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. CRC Press, Boca Raton.

Propriedade de Taylor e curtose em modelos MA

Esmeralda Gonçalves

Dep. de Matemática da Univ. Coimbra e CMUC, *esmerald@mat.uc.pt*

Cristina Martins

Dep. de Matemática da Univ. Coimbra, *cmtm@mat.uc.pt*

Nazaré Mendes-Lopes

Dep. de Matemática da Univ. Coimbra e CMUC, *nazare@mat.uc.pt*

Palavras-chave: Modelos MA; Curtose; Propriedade de Taylor.

Resumo: Neste trabalho estuda-se a ocorrência da propriedade de Taylor numa classe de modelos médias móveis não negativos, obtendo-se uma condição necessária e suficiente para que tais modelos possuam a referida propriedade. Os resultados obtidos são analisados em modelos com características de curtose significativamente diferentes, permitindo tirar conclusões sobre a relação entre a presença da propriedade de Taylor e a curtose do modelo.

1 Introdução

A presença da propriedade de Taylor em modelos de séries temporais tem despertado o interesse de vários autores desde que Taylor ([6]) constatou, na análise de várias séries financeiras, que as autocorrelações das observações em valor absoluto eram sistematicamente superiores às autocorrelações da mesma ordem para as observações ao quadrado. Aquela propriedade é a tradução teórica desta característica empírica, denominada efeito de Taylor.

Para cada $h \in \mathbb{Z}$ tal que $\rho_{|X|}(h)$ e $\rho_{X^2}(h)$ são estritamente positivas,

a propriedade de Taylor de ordem h é expressa pela condição

$$\rho_{|X|}(h) > \rho_{X^2}(h). \quad (1)$$

Reconhecido o efeito de Taylor como um facto estilizado presente em certo tipo de dados temporais, torna-se evidente o interesse de encontrar, nas diferentes classes de modelos dedicados a séries temporais, condições mediante as quais a propriedade de Taylor seja válida.

A complexidade teórica deste estudo, resultante em particular da necessidade de conhecer e comparar momentos de funções do processo, leva a que os resultados existentes sobre o tema sejam limitados e restritos a modelos de ordens baixas.

Por outro lado, em consonância com o facto do efeito de Taylor ter sido inicialmente observado em séries financeiras, a correspondente propriedade começou por ser analisada em modelos adequados para este tipo de séries, como os condicionalmente heteroscedásticos, tendo-se mesmo começado por associar esta propriedade à presença de volatilidade no modelo. São de referir neste caso os trabalhos de He e Teräsvirta ([5]), Gonçalves, Leite e Mendes-Lopes ([1]), Haas ([4]).

Os resultados obtidos por estes autores revelaram no entanto que a presença desta propriedade estava sobretudo relacionada com os valores elevados da curtose do modelo. Assim, entendeu-se relevante alargar o estudo relativo à presença da propriedade de Taylor a outras classes de modelos de séries temporais. Neste sentido, Gonçalves, Martins e Mendes-Lopes ([2],[3]) investigaram a presença da propriedade em modelos bilineares e em modelos autorregressivos não negativos de ordem 1, tendo novamente constatado a relevância da leptocurtose do modelo na sua ocorrência.

Dando continuidade a este estudo, temos como objetivo neste trabalho avaliar a presença da propriedade de Taylor em processos médias móveis (MA) não negativos de ordem 1, procurando identificar as características dos modelos que determinam tal presença. Assim começamos por estabelecer uma condição necessária e suficiente para

a presença da propriedade de Taylor no modelo MA(1). Prosseguiamos com aplicações deste estudo a modelos MA com processos de erro com características distribucionais significativamente distintas, sendo mais uma vez notória a importância da curtose na presença da referida propriedade no modelo.

2 Modelo MA(1) e propriedade de Taylor

Seja $X = (X_t, t \in \mathbb{Z})$ um processo estocástico real tal que

$$X_t = \phi \varepsilon_{t-1} + \varepsilon_t \quad (2)$$

onde ϕ é uma constante não negativa e $\varepsilon = (\varepsilon_t, t \in \mathbb{Z})$ é um processo estocástico não negativo, de componentes identicamente distribuídas admitindo momento de ordem 4, m_4 . Além disso, sendo \mathcal{E}_{t-1} a σ -álgebra gerada por $\varepsilon_{t-1}, \varepsilon_{t-2}, \dots$, supõe-se que, para $i = 1, 2, 3$, $E(\varepsilon_t^i | \mathcal{E}_{t-1})$ é independente de t , usando-se a notação $m_i = E(\varepsilon_t^i | \mathcal{E}_{t-1})$.

No lema seguinte apresentam-se as funções de autocorrelação dos processos X e X^2 .

Lema 2.1 *Tem-se*

a) $\rho_X(1) = \frac{\phi}{1+\phi^2}$

b) $\rho_{X^2}(1) = \frac{\phi}{V(X_t^2)} [\phi (V(\varepsilon_t^2) + 4m_1^2 V(\varepsilon_t)) + 2m_1(1 + \phi^2) Cov(\varepsilon_t, \varepsilon_t^2)]$

c) $\rho_X(h) = \rho_{X^2}(h) = 0$, para $h > 1$.

Dem.: Demonstram-se apenas os resultados correspondentes ao processo X^2 . No caso do processo X , o procedimento é semelhante.

De (2) obtém-se $X_t^2 = \phi^2 \varepsilon_{t-1}^2 + \varepsilon_t^2 + 2\phi \varepsilon_t \varepsilon_{t-1}$, pelo que, para $h \in \mathbb{N}$,

$$\begin{aligned} Cov(X_t^2, X_{t-h}^2) &= \phi^2 Cov(\varepsilon_{t-1}^2, X_{t-h}^2) + Cov(\varepsilon_t^2, X_{t-h}^2) \\ &\quad + 2\phi Cov(\varepsilon_t \varepsilon_{t-1}, X_{t-h}^2). \end{aligned}$$

Mas $Cov(\varepsilon_t^2, X_{t-h}^2) = 0$, uma vez que

$$\begin{aligned} Cov(\varepsilon_t^2, \varepsilon_{t-h-1}^2) &= E[\varepsilon_{t-h-1}^2 E(\varepsilon_t^2 | \varepsilon_{t-1})] - m_2^2 \\ &= m_2 E(\varepsilon_{t-h-1}^2) - m_2^2 = 0, \end{aligned}$$

o mesmo se verificando, com argumentos análogos, para $Cov(\varepsilon_t^2, \varepsilon_{t-h}^2)$ e $Cov(\varepsilon_t^2, \varepsilon_{t-h}\varepsilon_{t-h-1})$.

Para $h > 1$, usando novamente as propriedades da esperança condicionada, obtém-se ainda $Cov(\varepsilon_{t-1}^2, X_{t-h}^2) = Cov(\varepsilon_t \varepsilon_{t-1}, X_{t-h}^2) = 0$. Fica assim provada a parte c) do lema.

Consideremos agora $h = 1$, caso em que se tem

$$\begin{aligned} Cov(\varepsilon_t^2, X_t^2) &= \phi^2 Cov(\varepsilon_t^2, \varepsilon_{t-1}^2) + V(\varepsilon_t^2) + 2\phi Cov(\varepsilon_t^2, \varepsilon_t \varepsilon_{t-1}) \\ &= V(\varepsilon_t^2) + 2\phi m_1 m_3 - m_1^2 m_2 \\ &= V(\varepsilon_t^2) + 2\phi m_1 Cov(\varepsilon_t, \varepsilon_t^2) \end{aligned}$$

$$\text{e } Cov(\varepsilon_t \varepsilon_{t-1}, X_{t-1}^2) = 2\phi m_1^2 V(\varepsilon_t) + m_1 Cov(\varepsilon_t, \varepsilon_t^2).$$

A partir destes dois resultados conclui-se então que

$$Cov(X_t^2, X_{t-1}^2) = \phi^2 [V(\varepsilon_t^2) + 4m_1^2 V(\varepsilon_t)] + 2\phi(1 + \phi^2)m_1 Cov(\varepsilon_t, \varepsilon_t^2),$$

o que conduz diretamente ao resultado enunciado na parte b) do lema. ■

Do presente lema decorre que, neste tipo de modelos, a propriedade de Taylor só poderá verificar-se para $h = 1$. Nota-se ainda que os valores $\rho_X(1)$ e $\rho_{X^2}(1)$ são positivos, atendendo em particular ao facto de $Cov(\varepsilon_t, \varepsilon_t^2) \geq 0$, uma vez que ε_t é não negativo. Como consequência imediata deste lema temos então o seguinte teorema:

Teorema 2.2 *O processo X definido em (2) verifica a propriedade de Taylor (1) se e só se $h = 1$ e*

$$\frac{V(X_t^2)}{1 + \phi^2} > \phi [V(\varepsilon_t^2) + 4m_1^2 V(\varepsilon_t)] + 2m_1(1 + \phi^2) Cov(\varepsilon_t, \varepsilon_t^2). \quad (3)$$

A condição (3) relaciona momentos do processo X com momentos do correspondente processo de erro; notamos que de (2) e atendendo às propriedades da esperança condicionada, obtém-se a expressão que se segue para os momentos de X em função dos momentos de ε .

$$E(X_t^n) = (1 + \phi^n)m_n + \sum_{i=1}^{n-1} \binom{n}{i} \phi^i m_i m_{n-i}. \quad (4)$$

3 Propriedade de Taylor e curtose

Nesta secção avaliamos a presença da propriedade de Taylor no modelo (2) considerando processos de erro com distribuições não negativas e com pesos nas caudas significativamente diferentes.

Com o objetivo de relacionar as curtoses dos processos X e ε , começamos por exprimir os momentos centrados de ordem k de X , $\mu_{k,X}$, $k = 2, 4$, em termos dos momentos centrados de ε , $\mu_{k,\varepsilon}$, $k = 2, 4$.

De (4), obtemos $E(X_t) = (1 + \phi)m_1$ e $E(X_t^2) = (1 + \phi^2)m_2 + 2\phi m_1^2$, resultando $\mu_{2,X} = (1 + \phi^2)\mu_{2,\varepsilon}$. De (2) e tendo em conta a expressão de $E(X_t)$, podemos escrever

$$\mu_{4,X} = E\left[\left((\varepsilon_t - m_1) + \phi(\varepsilon_{t-1} - m_1)\right)^4\right] = (1 + \phi^4)\mu_{4,\varepsilon} + 6\phi^2\mu_{2,\varepsilon}^2.$$

Conclui-se então que a curtose de X , K_X , e a curtose de ε , K_ε , estão relacionadas da seguinte forma:

$$K_X(\phi) = \frac{6\phi^2}{(1 + \phi^2)^2} + \frac{1 + \phi^4}{(1 + \phi^2)^2} K_\varepsilon. \quad (5)$$

Desta igualdade deduz-se facilmente que o processo X é leptocúrtico, platicúrtico ou mesocúrtico se e só se o processo ε é, respetivamente, leptocúrtico, platicúrtico ou mesocúrtico.

Da igualdade (5) conclui-se também que $K_X(\phi)$ tende para K_ε , tanto no caso de ϕ tender para 0 como no caso de ϕ tender para $+\infty$.

Deduz-se ainda que, se $K_\varepsilon < 3$ então $K_X(\phi)$ é crescente para $0 < \phi < 1$ e decrescente para $\phi > 1$, atingindo no ponto 1 o seu valor máximo, $\frac{1}{2}(K_\varepsilon + 3)$. Por outro lado, se $K_\varepsilon > 3$ então $K_X(\phi)$

é decrescente para $0 < \phi < 1$ e crescente para $\phi > 1$, atingindo no ponto 1 o seu valor mínimo, $\frac{1}{2}(K_\varepsilon + 3)$.

No estudo que se segue, a igualdade (3) será usada na forma $T_{L_\varepsilon}(\phi) > 0$, com

$$T_{L_\varepsilon}(\phi) = \frac{V(X_t^2)}{1 + \phi^2} - \phi [V(\varepsilon_t^2) + 4m_1^2 V(\varepsilon_t)] - 2m_1(1 + \phi^2) \text{Cov}(\varepsilon_t, \varepsilon_t^2), \quad (6)$$

onde L_ε designa a lei marginal do processo de erro. Apresentam-se em apêndice as expressões de $T_{L_\varepsilon}(\phi)$ para as leis L_ε consideradas nas secções seguintes.

3.1 Erros com distribuição leptocúrtica

Suponhamos que ε_t segue a lei gama de parâmetros α e θ , $\gamma(\alpha, \theta)$, $\alpha > 0$, $\theta > 0$, cuja curtose é dada por $K_\varepsilon = 3 + \frac{6}{\theta}$.

Verifica-se que tanto $T_{\gamma(\alpha, \theta)}$ como K_X são funções de ϕ e θ , mas não de α . Na Figura 1 apresentam-se os gráficos de $K = K_X(\phi)$ e $T = T_{\gamma(\alpha, \theta)}(\phi)$, para $0 < \phi < 5$, $0 < \theta < 10$. Observa-se que $T_{\gamma(\alpha, \theta)}(\phi) > 0$, ou seja, a propriedade de Taylor está sempre presente e, de um modo geral, é tanto mais acentuada quanto maior é a curtose de X . Nota-se que, com o aumento de θ , o valor de K diminui o mesmo acontecendo com o valor de T . Concretamente, quando θ tende para $+\infty$, tem-se K a tender para 3 e T a tender para zero, como se pode constatar pela expressão desta função, (A.1), apresentada em apêndice. Este resultado vai ao encontro do que tem sido observado noutros casos já estudados ([2],[3],[4]): a propriedade de Taylor manifesta-se, em geral, de modo mais acentuado para maiores valores da curtose do processo.

Obtêm-se conclusões análogas quando se considera para ε_t a distribuição de Pareto de parâmetros α e θ , $Par(\alpha, \theta)$, $\alpha > 0$, $\theta > 4$ (para assegurar a existência do momento de ordem 4), cuja curtose é dada por $K_\varepsilon = 3 + \frac{6(\theta^3 + \theta^2 - 6\theta - 2)}{\theta(\theta - 3)(\theta - 4)}$. Neste caso, K_X e $T_{Par(\alpha, \theta)}$ também dependem de ϕ e θ mas não de α .

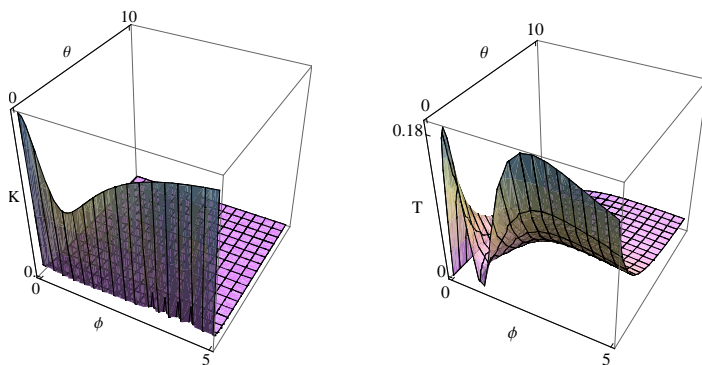


Figura 1: Gráficos de $K = K_X(\phi)$ com $\varepsilon_t \sim \gamma(\alpha, \theta)$ (esq.) e de $T = T_{\gamma(\alpha, \theta)}(\phi)$ (dir.), $0 < \phi < 5$, $0 < \theta < 10$

Com o objetivo de comparar $K_X(\phi)$ nos casos Pareto e gama bem como $T_{Par(\alpha, \theta)}(\phi)$ com $T_{\gamma(\alpha, \theta)}(\phi)$, apresenta-se, na Figura 2, o gráfico da diferença entre as curtoses de X nos casos Pareto e gama e o gráfico da diferença $T_{Par(\alpha, \theta)}(\phi) - T_{\gamma(\alpha, \theta)}(\phi)$, ambos para $0 < \phi < 3$ e $4 < \theta < 6$. Observa-se que o processo X tem maior curtose no caso Pareto e, de um modo geral, a propriedade de Taylor surge com maior intensidade nesse caso. Nos casos em que a diferença $T_{Par(\alpha, \theta)}(\phi) - T_{\gamma(\alpha, \theta)}(\phi)$ assume valores negativos, tais valores são próximos de 0 e ocorrem nas vizinhanças de $\phi = 1$, caso em que a curtose de ambos os modelos é mínima.

3.2 Erros com distribuição platicúrtica

Nos casos leptocúrticos analisados, a propriedade de Taylor está sempre presente. No entanto, esta propriedade também se manifesta em modelos MA(1) platicúrticos, embora de forma menos acentuada, como mostram os exemplos seguintes.

Consideremos o processo X definido em (2) com ε_t seguindo a lei beta de parâmetros α e θ , $B(\alpha, \theta)$, $\alpha > 0$, $\theta > 0$. Mais precisamente,

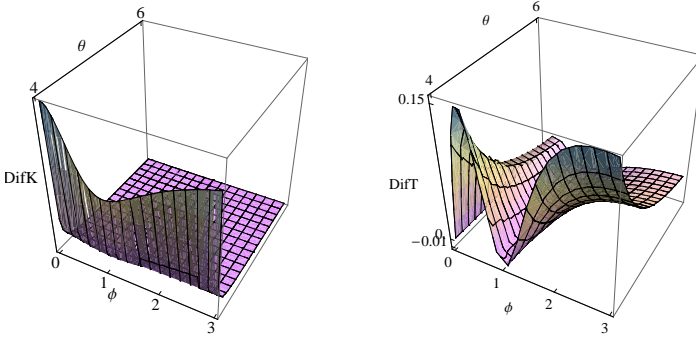


Figura 2: Gráfico da diferença, $DifK$, entre $K_X(\phi)$ no caso Pareto e $K_X(\phi)$ no caso gama (esq.) e gráfico de $DifT = T_{Par(\alpha, \theta)}(\phi) - T_{\gamma(\alpha, \theta)}(\phi)$ (dir.), $0 < \phi < 3$, $4 < \theta < 6$

analisamos os casos $\theta = \frac{\alpha}{2}$ (distribuição assimétrica negativa), $\theta = \alpha$ (distribuição simétrica) e $\theta = 2\alpha$ (distribuição assimétrica positiva). Relativamente aos valores da curtose, tem-se $K_\varepsilon = 3 - \frac{6}{4+3\alpha}$ quando $\theta = \frac{\alpha}{2}$, $K_\varepsilon = 3 - \frac{6}{3+2\alpha}$ quando $\theta = \alpha$ e $K_\varepsilon = 3 - \frac{6}{2+3\alpha}$ quando $\theta = 2\alpha$.

Em todos estes casos, a curtose de X e a função T_{L_ε} dependem de ϕ e de α , tendo-se $T_{B(\alpha, \alpha)}(\phi) > 0$ e $T_{B(\alpha, 2\alpha)}(\phi) > 0$, $\phi > 0$, $\alpha > 0$, como facilmente se deduz das expressões destas funções, (A.3) e (A.4), apresentadas em apêndice.

Na Figura 3 apresentam-se, para os casos $B(\alpha, \alpha)$ e $B(\alpha, 2\alpha)$, os gráficos sobrepostos tanto das curtoses como das funções T_{L_ε} para os correspondentes processos MA(1).

Observa-se que, embora presente, a propriedade de Taylor manifesta-se de forma ténue (isto é, T_{L_ε} assume valores próximos de zero). De notar ainda que, apesar da fraca presença, a propriedade manifesta-se de forma mais acentuada no caso da distribuição $B(\alpha, 2\alpha)$, que é, dos dois casos considerados, aquele em que a curtose do processo X é maior. Observa-se ainda que, fixando ϕ , o aumento de α acarreta um aumento da curtose de X , mas uma presença mais ténue da

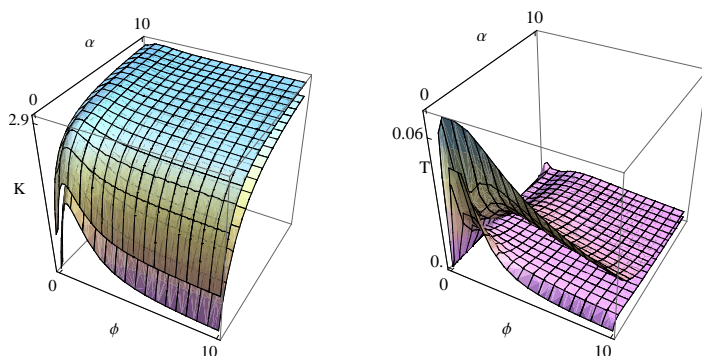


Figura 3: Gráficos de $K = K_X(\phi)$ nos casos $B(\alpha, 2\alpha)$ (por cima) e $B(\alpha, \alpha)$ (por baixo), à esquerda, e gráficos de $T = T_{B(\alpha, 2\alpha)}(\phi)$ (por cima) e de $T = T_{B(\alpha, \alpha)}(\phi)$ (por baixo), à direita, $0 < \phi < 10$, $0 < \alpha < 10$

propriedade de Taylor.

No caso assimétrico negativo $B(\alpha, \frac{\alpha}{2})$, o eixo das ordenadas da Figura 4 mostra que a propriedade de Taylor nem sempre está presente e, quando está, a sua presença manifesta-se de forma ténue tal como nos outros casos platicúrticos analisados.

4 Conclusão

Neste trabalho constatamos que, tal como acontece noutros modelos lineares ou não lineares estudados, continua a ser notória a existência de uma forte ligação entre a propriedade de Taylor e a curtose do processo. Efetivamente, nos modelos leptocúrticos analisados a propriedade de Taylor está sempre presente e tal presença acentua-se, em geral, com o aumento da curtose do processo. Relativamente aos casos platicúrticos em que a propriedade de Taylor está presente, ela manifesta-se de modo suave.

Verificou-se ainda que, nos exemplos correspondentes a modelos em que o processo gerador tem distribuição simétrica ou assimétrica

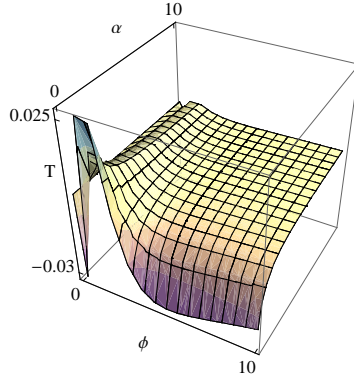


Figura 4: Gráfico de $T = T_{B(\alpha, \frac{\alpha}{2})}(\phi)$, $0 < \phi < 10$, $0 < \alpha < 10$

positiva, a propriedade de Taylor está sempre presente. No modelo em que o processo gerador tem distribuição assimétrica negativa, a propriedade nem sempre está presente. Foram analisados outros casos nestas condições tendo-se verificado o mesmo resultado.

Uma extensão natural do presente trabalho consistirá no desenvolvimento de estudos teóricos que, com base nesta análise prospetiva, permitam explicar a presença da propriedade de Taylor em modelos com tais características distribucionais, designadamente leptocurtose, simetria ou assimetria positiva.

Outro aspeto relevante será o de alargar este estudo a modelos de qualquer sinal sendo, neste caso, necessário ultrapassar o problema, não trivial, do estudo da função de autocorrelação do módulo do processo. Algumas análises empíricas levam-nos a conjecturar que neste caso a propriedade deverá traduzir-se por

$$|\rho_{|X|}(h)| > |\rho_{X^2}(h)|, \quad h \in \mathbb{Z}.$$

A consideração de modelos dos tipos já estudados, mas de ordens superiores, é também uma questão em aberto e sobre a qual se esperam futuros desenvolvimentos.

Apêndice - Expressões de $T_{L_\varepsilon}(\phi)$ para as leis L_ε consideradas

$$T_{\gamma(\alpha,\theta)}(\phi) = \frac{\phi}{1+\phi^2} \frac{-2\phi^2\theta+3(1+\theta)(1+\phi^4)-\phi(3+\theta)(1+\phi^2)}{4\phi\theta(1+\theta)(1+\phi^2)+2\phi^2\theta(1+2\theta)+(1+\phi^4)(3+5\theta+2\theta^2)} \quad (\text{A.1})$$

$$T_{Par(\alpha,\theta)}(\phi) = \frac{\phi}{1+\phi^2} \frac{N_{Par(\alpha,\theta)}(\phi)}{D_{Par(\alpha,\theta)}(\phi)}, \quad (\text{A.2})$$

com

$$\begin{aligned} N_{Par(\alpha,\theta)}(\phi) &= \phi(1+\phi^2)(3+3\theta+2\theta^2-2\theta^3) \\ &\quad -\phi^2\theta(4+11\theta-11\theta^2+2\theta^3) \\ &\quad +(1+\phi^4)(-3+5\theta-3\theta^3+\theta^4), \\ D_{Par(\alpha,\theta)}(\phi) &= (1+\phi^4)(\theta-3)(\theta-1)^4 \\ &\quad +2\phi(1+\phi^2)\theta(\theta-1)^2(8-6\theta+\theta^2) \\ &\quad +\phi^2\theta(12-55\theta+53\theta^2-18\theta^3+2\theta^4) \end{aligned}$$

$$T_{B(\alpha,2\alpha)}(\phi) = \frac{\phi}{1+\phi^2} \frac{3-\phi+4\phi^2-\phi^3+3\phi^4+6\alpha(2-\phi-\phi^3+\phi^4)}{D_{B(\alpha,2\alpha)}(\phi)}, \quad (\text{A.3})$$

com

$$\begin{aligned} D_{B(\alpha,2\alpha)}(\phi) &= 9+12\phi+16\phi^2+12\phi^3+9\phi^4 \\ &\quad +18\alpha^2(1+\phi)^2(1+\phi^2) \\ &\quad +12\alpha(3+4\phi+4\phi^2+4\phi^3+3\phi^4) \end{aligned}$$

$$T_{B(\alpha,\alpha)}(\phi) = \frac{\phi}{1+\phi^2} \frac{3\phi^2+\alpha(1+\phi^2)(1-\phi+\phi^2)}{D_{B(\alpha,\alpha)}(\phi)}, \quad (\text{A.4})$$

com

$$\begin{aligned} D_{B(\alpha,\alpha)}(\phi) &= 4\alpha^2(1+\phi)^2(1+\phi^2)+3(1+\phi+\phi^2)^2 \\ &\quad +\alpha(9+16\phi+18\phi^2+16\phi^3+9\phi^4) \end{aligned}$$

$$T_{B(\alpha,\frac{\alpha}{2})}(\phi) = -\frac{\phi}{1+\phi^2} \frac{3+\phi-16\phi^2+\phi^3+3\phi^4+3\alpha(1+\phi-6\phi^2+\phi^3+\phi^4)}{D_{B(\alpha,\frac{\alpha}{2})}(\phi)}, \quad (\text{A.5})$$

com

$$\begin{aligned} D_{B(\alpha,\frac{\alpha}{2})}(\phi) &= 9+24\phi+40\phi^2+24\phi^3+9\phi^4 \\ &\quad +18\alpha^2(1+\phi)^2(1+\phi^2) \\ &\quad +3\alpha(9+20\phi+26\phi^2+20\phi^3+9\phi^4) \end{aligned}$$

Referências

- [1] Gonçalves, E., Leite, J., Mendes-Lopes, N. (2009). A mathematical approach to detect the Taylor property in TARCH processes. *Statist. Probab. Lett.* 79, 602–610.
- [2] Gonçalves, E., Martins, C.M., Mendes-Lopes, N. (2015). The Taylor property in bilinear models. *RevStat - Statistical Journal* 13, 3, 207–226.
- [3] Gonçalves, E., Martins, C.M., Mendes-Lopes, N. (2014). Propriedade de Taylor em processos autorregressivos. In Pereira, I., Freitas, A., Scotto, M., Silva, M.E., Paulino, C.D. (eds.): *Estatística: A ciência da incerteza, Atas do XXI Congresso Anual da Sociedade Portuguesa de Estatística* 51–53, Edições SPE.
- [4] Haas, M. (2009). Persistence in volatility, conditional kurtosis, and the Taylor property in absolute value GARCH processes. *Statist. Probab. Lett.* 79, 5, 1674–1683.
- [5] He, C., Teräsvirta, H. (1999). Properties of moments of a family of GARCH processes. *J. Econom.* 92, 173–192.
- [6] Taylor, S. (1986). *Modelling Financial Time Series*. Wiley.

Números de clientes servidos e bloqueados em períodos de ocupação contínua de filas $M/M/1/n$ com bloqueio

Fátima Ferreira

Universidade de Trás-os-Montes e Alto Douro, UTAD, Departamento de Matemática e CEMAT, *mmferrei@utad.pt*

António Pacheco

Instituto Superior Técnico, Universidade de Lisboa, Departamento de Matemática e CEMAT, *apacheco@math.tecnico.ulisboa.pt*

Helena Ribeiro

Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria e CEMAT, *helena.ribeiro@ipleiria.pt*

Palavras-chave: Filas de espera; Períodos de ocupação contínua; Uniformização; Cadeias de Markov.

Resumo: Neste trabalho analisamos filas de espera $M/M/1/n$ com bloqueio. Nestes sistemas, a admissão dos clientes é modulada pelo estado do sistema nos instantes de chegada, *i.e.*, consoante a dimensão da fila aquando da sua chegada, os clientes decidem com determinada probabilidade entrar no sistema. Tirando partido da propriedade regenerativa markoviana da cadeia embebida nos instantes de chegada ou saída de clientes, caracterizamos a distribuição de probabilidade conjunta do número de clientes servidos e do número de clientes perdidos em períodos de ocupação contínua. Apresentamos um algoritmo recursivo para o cálculo da respetiva função geradora de probabilidades. Terminamos com uma breve ilustração numérica dos resultados derivados.

1 Introdução

Neste trabalho estudamos filas de espera markovianas de capacidade finita com um único servidor e com diferentes políticas de bloqueio de clientes à chegada dos mesmos ao sistema; na terminologia anglo-saxónica e usando a notação introduzida por David G. Kendall, tal corresponde a uma fila de espera $M/M/1/n$ with *balking*. Filas de espera com bloqueio possuem muitas aplicações porque é frequente os clientes terem a possibilidade de postergar ou desistir de um dado tipo de serviço quando, à chegada ao sistema, o nível de congestionamento do servidor é considerado insatisfatório por parte dos mesmos clientes. Com origem no trabalho de Haight [4], o estudo de filas com bloqueio (*balking* e *reversed balking*) tem despertado o interesse de diversos autores (*c.f.* [1], [2], [5] e [6] e suas referências).

A análise de filas de espera com bloqueio de clientes em períodos de ocupação contínua, i.e., em períodos contínuos de utilização efetiva do servidor, é relevante do ponto de vista do operador e fornece informação crucial para a sua gestão. Consideram-se neste estudo períodos de ocupação contínua iniciados com múltiplos clientes no sistema. Em particular, um i -período de ocupação contínua (i -p.o.c.) representa um período que se inicia com i clientes no sistema, com um cliente a iniciar serviço nesse instante, e termina no instante subsequente em que o sistema fica vazio.

Este trabalho visa calcular a distribuição de probabilidade conjunta do número de clientes servidos e do número de clientes perdidos em períodos de ocupação contínua. Nesse âmbito, na Secção 2, caracterizamos o sistema e mostramos como combinar as cadeias embebidas com a uniformização por forma a calcular a matriz de probabilidades de transição da cadeia embebida nos instantes de chegadas e saídas de clientes no sistema. Na Secção 3, tirando partido da estrutura markoviana de filas de espera $M/M/1/n$ com bloqueio, obtém-se a função geradora de probabilidades e a respetiva distribuição conjunta do número de clientes servidos e do número de clientes perdidos em períodos de ocupação contínua nestes sistemas. Finalmente na Secção 4 ilustramos os resultados obtidos.

2 Caracterização do sistema e do sistema embebido

As filas de espera $M/M/1/n$ com bloqueio são sistemas de capacidade finita, n , às quais os clientes chegam segundo um processo de Poisson de taxa λ e são servidos à taxa μ por um único servidor. Os tempos de serviço e os tempos entre chegadas de clientes são variáveis aleatórias independentes com distribuição exponencial. Se um cliente encontra à chegada n clientes no sistema, ele é bloqueado com probabilidade 1. O que distingue estas filas das $M/M/1/n$ tradicionais é o facto de a admissão dos clientes ser modulada pelo estado do sistema nos instantes de chegada. Especificamente, se à chegada um cliente encontra i clientes no sistema, ainda que tenha lugar na fila ($i < n$), ele decide não entrar com probabilidade $1 - e_i$. Seja $Y = (Y(t))_{t \in \mathbb{R}_0^+}$ o processo em tempo contínuo onde $Y(t)$ denota o número de clientes no instante t num sistema $M/M/1/n$ com bloqueio. Sejam $(\xi_n)_{n \in \mathbb{N}}$ a sucessão dos instantes de chegada de clientes, $(\varpi_n)_{n \in \mathbb{N}}$ a sucessão dos instantes de saída de clientes e $(\tau_n)_{n \in \mathbb{N}}$ a sucessão dos instantes de chegada ou saída de clientes no sistema. Dado que as sucessões $(\xi_{n+1} - \xi_n)_{n \in \mathbb{N}}$ e $(\varpi_{n+1} - \varpi_n)_{n \in \mathbb{N}}$ são independentes com $\xi_{n+1} - \xi_n \sim \exp(\lambda)$ e $\varpi_{n+1} - \varpi_n \sim \exp(\mu)$. Y é um processo regenerativo Markoviano (*c.f.* [3]) associado à sequência de renovoamento $(\tau_n)_{n \in \mathbb{N}}$, com espaço de estados $E = \{0, 1, \dots, n\}$ e matriz geradora infinitesimal $Q = (q_{ij})_{i,j \in E}$,

$$q_{ij} = \begin{cases} \lambda e_i, & j = i + 1 \wedge i \neq n \\ -\lambda e_0, & j = i = 0 \\ -(\lambda e_i + \mu), & j = i \neq 0 \\ \mu, & j = i - 1 \wedge i \neq 0 \\ 0, & j \in E \setminus \{i - 1, i, i + 1\} \end{cases}.$$

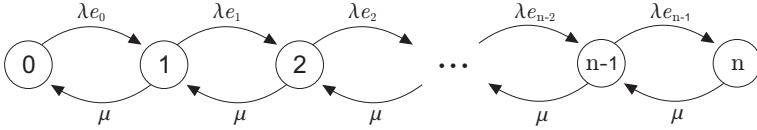


Figura 1: Diagrama de transição do sistema $M/M/1/n$ com bloqueio

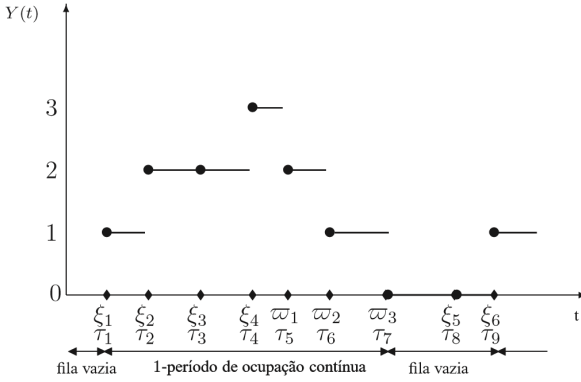


Figura 2: Trajetória típica no sistema $M/M/1/n$ com bloqueio.

Nas Figuras 1 e 2 apresentamos, respetivamente, o diagrama de transição do sistema $M/M/1/n$ com bloqueio e uma trajetória da evolução do número de clientes no sistema ao longo do tempo.

Como a taxa de saída dos estados é majorada por uma constante finita, $\max |q_{ii}| \leq \lambda + \mu < \infty$, consideramos a cadeia de Markov em tempo discreto embebida em $(\tau_k)_{k \in \mathbb{N}}$, uniformizada à taxa $\lambda + \mu$, $\bar{Y} = (\bar{Y}_k)_{k \in \mathbb{N}}$, onde \bar{Y}_k denota o número de clientes no sistema no instante τ_k . A matriz de probabilidades de transição a um passo de \bar{Y} , $P = (p_{ij})_{i,j \in E}$, é tal que $P = I + \frac{Q}{\mu + \lambda}$ onde, para $i, j \in E$,

$$p_{ij} = \begin{cases} \frac{\lambda e_i}{\lambda + \mu}, & j = i + 1 \wedge i \neq n \\ \frac{\mu + \lambda(1 - e_0)}{\lambda + \mu}, & j = i = 0 \\ \frac{\lambda(1 - e_i)}{\lambda + \mu}, & j = i \neq 0 \\ \frac{\mu}{\lambda + \mu}, & j = i - 1 \wedge i \neq 0 \\ 0, & j \in E \setminus \{i - 1, i, i + 1\} \end{cases} . \quad (1)$$

3 Probabilidade conjunta de (S_i, L_i)

Nesta seção derivamos a distribuição de probabilidade conjunta do vetor aleatório (S_i, L_i) onde, para $i \in E \setminus \{0\}$, S_i denota o número de clientes servidos durante um i -p.o.c. e L_i denota o número de clientes perdidos durante um i -p.o.c.

Para o efeito, calculamos a função geradora de probabilidades de (S_i, L_i) :

$$g_i(u, v) = E(u^{S_i} v^{L_i}) = \sum_{s \in \mathbb{N}} \sum_{l \in \mathbb{N}_0} u^s v^l P(S_i = s, L_i = l) \quad (2)$$

com $|u| \leq 1$ e $|v| \leq 1$ donde, por derivação, obtemos a respetiva distribuição de probabilidade de (S_i, L_i) ,

$$P(S_i = s, L_i = 0) = \frac{1}{s!} \left. \frac{\partial^s g_i(u, 0)}{\partial u^s} \right|_{u=0}, \quad s \in \mathbb{N}$$

e

$$P(S_i = s, L_i = l) = \frac{1}{s! l!} \left. \frac{\partial^{s+l} g_i(u, v)}{\partial u^s \partial v^l} \right|_{u=0, v=0}, \quad (s, l) \in \mathbb{N} \times \mathbb{N}.$$

Condicionando no tipo de evento X que ocorre no instante τ_k da primeira transição após iniciar um i -p.o.c., com

$$X = \begin{cases} -1, & \text{se em } \tau_k \text{ ocorre a saída de um cliente} \\ 0, & \text{se em } \tau_k \text{ ocorre a chegada de um cliente sem entrada} \\ 1, & \text{se em } \tau_k \text{ ocorre a chegada de um cliente com entrada} \end{cases}$$

tem-se,

$$(S_n, L_n)|_{X=x} =_{st} \begin{cases} (1 + S_{n-1}, L_{n-1}), & x = -1 \\ (S_n, 1 + L_n), & x = 0 \end{cases}$$

e, para $i \in E \setminus \{0, n\}$,

$$(S_i, L_i)|_{X=x} =_{st} \begin{cases} (1 + S_{i-1}, L_{i-1}), & x = -1 \\ (S_i, 1 + L_i), & x = 0 \\ (S_{i+1}, L_{i+1}), & x = 1 \end{cases}.$$

Pelo teorema da probabilidade total, para $(s, l) \in \mathbb{N} \times \mathbb{N}_0$,

$$\begin{aligned} P(S_n = s, L_n = l) &= P(S_{n-1} = s - 1, L_{n-1} = l)P(X = -1) \\ &\quad + P(S_n = s, L_n = l - 1)P(X = 0) \end{aligned}$$

e, para $i \in E \setminus \{0, n\}$,

$$\begin{aligned} P(S_i = s, L_i = l) &= \mathbf{1}_{\{(i,s,l)=(1,1,0)\}}P(X = -1) \\ &\quad + \mathbf{1}_{\{i>1\}}P(S_{i-1} = s - 1, L_{i-1} = l)P(X = -1) \\ &\quad + P(S_i = s, L_i = l - 1)P(X = 0) \\ &\quad + P(S_{i+1} = s, L_{i+1} = l)P(X = 1) \end{aligned}$$

com $\mathbf{1}_{\{z\}}$ a denotar a função indicadora da condição z . Atendendo a (1) e (2) obtém-se:

$$g_n(u, v) = \frac{u\mu}{\lambda + \mu}g_{n-1}(u, v) + \frac{v\lambda}{\lambda + \mu}g_n(u, v)$$

e

$$g_i(u, v) = \frac{u\mu}{\lambda + \mu}g_{i-1}(u, v) + \frac{v\lambda(1 - e_i)}{\lambda + \mu}g_i(u, v) + \frac{\lambda e_i}{\lambda + \mu}g_{i+1}(u, v).$$

Assim,

$$g_i(u, v) = \theta_n g_{n-1}(u, v), \text{ com } \theta_n = \frac{u\mu}{\lambda + \mu - v\lambda} \quad (3)$$

e, para $i \in E \setminus \{0, n\}$,

$$g_i(u, v) = \frac{u\mu}{\lambda + \mu - v\lambda(1 - e_i)} g_{i-1}(u, v) + \frac{\lambda e_i}{\lambda + \mu - v\lambda(1 - e_i)} g_{i+1}(u, v). \quad (4)$$

Procedendo de modo recursivo para $i = n-1, n-2, \dots, 1$, escrevemos a função geradora de probabilidades de (S_i, L_i) em função da função geradora de probabilidades de (S_{i-1}, L_{i-1}) ,

$$g_i(u, v) = \theta_i g_{i-1}(u, v)$$

com

$$\theta_i = \frac{u\mu}{\lambda + \mu - v\lambda(1 - e_i) - \lambda e_i \theta_{i+1}}.$$

Finalizamos esta secção apresentando, na Figura 3, um procedimento recursivo para obter a função geradora de probabilidades de (S_i, L_i) em sistemas $M/M/1/n$ com bloqueio, usando o facto de $g_0(u, v) = 1$.

4 Ilustração numérica

Para efeitos de ilustração numérica, consideramos sistemas $M/M/1/n$ com taxa de serviço unitária ($\mu = 1$) e três políticas de bloqueio de clientes:

- os clientes são bloqueados apenas quando o sistema está cheio,

$$e^{(1)} = (e_0, e_1, \dots, e_n) \text{ com } e_i = \begin{cases} 1 & i \in E \setminus \{n\} \\ 0 & i = n \end{cases}$$

que representa o bloqueio usual das filas $M/M/1/n$;

Figura 3: Algoritmo para calcular a função geradora de probabilidades de (S_i, L_i) em sistemas $M/M/1/n$ com bloqueio.

Input: $n, \lambda, \mu, e = (e_0, e_1, \dots, e_{n-1}, 1)$

$$\theta_n(u, v) = \frac{u\mu}{\lambda + \mu - v\lambda};$$

For $i = n - 1 : -1 : 1$

$$\theta_i(u, v) = \frac{u\mu}{\lambda + \mu - v\lambda(1 - e_i) - \lambda e_{i-1}\theta_{i+1}(u, v)};$$

End for

$$g_1(u, v) = \theta_1(u, v);$$

For $i = 2 : n$

$$g_i(u, v) = \theta_i(u, v) g_{i-1}(u, v);$$

End for

Output: $(g_i(u, v))_{i=1,2,\dots,n}$

- um bloqueio de clientes que aumenta com o tamanho da fila,

$$e^{(2)} = (e_0, e_1, \dots, e_n) \text{ com } e_i = \begin{cases} \frac{1}{i+2} & i \in E \setminus \{n\} \\ 0 & i = n \end{cases}$$

situação tradicional quando os clientes têm pressa de ser atendidos;

- um bloqueio reverso que privilegia a entrada de clientes quando o sistema está mais ocupado,

$$e^{(3)} = (e_0, e_1, \dots, e_n) \text{ com } e_i = \begin{cases} \frac{1}{n+1-i} & i \in E \setminus \{n\} \\ 0 & i = n \end{cases}$$

situação que pode ocorrer, por exemplo, na restauração ou em investimentos na bolsa.

As Tabelas 1 e 2 apresentam a função de probabilidade conjunta do número de clientes servidos e do número de clientes perdidos num período de ocupação contínua que inicia com um cliente no sistema em filas $M/M/1/7$ com política de bloqueio $e^{(2)} = (\frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{8}, 0)$ e taxas de chegada $\lambda = 0.5$ e $\lambda = 1.1$, respetivamente.

0.6667	0.1481	0.0329	0.0073	0.0016	0.0004	0.0001	0.0000	0.0000	0.0000
0.0494	0.0343	0.0159	0.0061	0.0021	0.0007	0.0002	0.0001	0.0000	0.0000
0.0064	0.0076	0.0054	0.0030	0.0014	0.0006	0.0002	0.0001	0.0000	0.0000
0.0010	0.0016	0.0015	0.0011	0.0007	0.0004	0.0002	0.0001	0.0000	0.0000
0.0001	0.0003	0.0004	0.0004	0.0003	0.0002	0.0001	0.0001	0.0000	0.0000
0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Tabela 1: Probabilidade conjunta do número de clientes servidos e do número de clientes perdidos num 1-p.o.c., no sistema $M/M/1/7$ com taxa de serviço $\mu = 1$ e taxa de chegadas $\lambda = 0.50$, $P(S_1 = s, L_1 = l)$ para $s = 0, 1, \dots, 9$ e $l = 0, 1, \dots, 10$.

No sistema com baixa intensidade de tráfego (Tabela 1), a taxa de entradas é bem menor do que a taxa de serviço pelo que a fila tem pouca tendência a encher e a ter valores elevados de perdas. Consequentemente a massa de probabilidade conjunta concentra-se nos valores mais baixos do número de clientes servidos e do número de clientes perdidos. Note-se que, neste caso, $P(S_1 \leq 3, L_1 \leq 3) = 0.9831$. Em contraste, no sistema com taxa de utilização mais elevada (Tabela 2), com mais tendência para encher e consequentemente para ocorrerem mais bloqueios, $P(S_1 \leq 3, L_1 \leq 3) = 0.8933$ dado que a massa de probabilidade conjunta ainda apresenta probabilidades positivas para valores mais elevados do número de clientes servidos e do número de clientes perdidos.

A Figura 4 ilustra a sensibilidade da distribuição conjunta acumulada de (S_1, L_1) no ponto (3,3) de filas $M/M/1/10$ em função da taxa de utilização ($\rho = \lambda$), para três diferentes políticas de bloqueio, $e^{(1)} = (1, 1, \dots, 1, 0)$, $e^{(2)} = (\frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{11}, 0)$ e $e^{(3)} = (\frac{1}{11}, \frac{1}{10}, \dots, \frac{1}{2}, 0)$.

0.4762	0.1663	0.0581	0.0203	0.0071	0.0025	0.0009	0.0003	0.0001	0.0000
0.0396	0.0432	0.0315	0.0191	0.0104	0.0053	0.0026	0.0012	0.0006	0.0003
0.0058	0.0107	0.0120	0.0105	0.0078	0.0053	0.0033	0.0019	0.0011	0.0006
0.0010	0.0025	0.0039	0.0044	0.0042	0.0035	0.0027	0.0019	0.0013	0.0008
0.0002	0.0006	0.0011	0.0016	0.0018	0.0018	0.0016	0.0014	0.0010	0.0008
0.0000	0.0001	0.0003	0.0005	0.0007	0.0008	0.0008	0.0008	0.0007	0.0006
0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0004	0.0004	0.0004	0.0003
0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001

Tabela 2: Probabilidade conjunta do número de clientes servidos e do número de clientes perdidos num 1-p.o.c., no sistema $M/M/1/7$ com taxa de serviço $\mu = 1$ e taxa de chegadas $\lambda = 1.10$, $P(S_1 = s, L_1 = l)$ para $s = 0, 1, \dots, 9$ e $l = 0, 1, \dots, 10$.

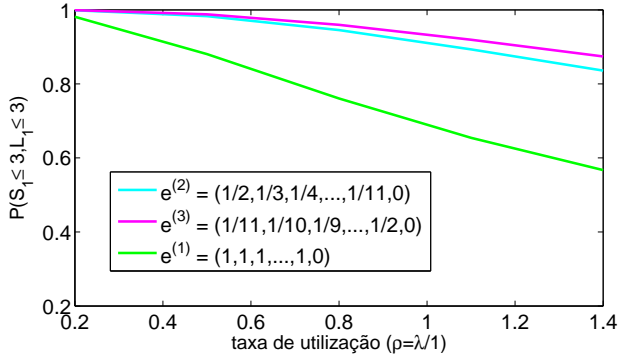


Figura 4: Probabilidade conjunta acumulada do número de clientes servidos e do número de clientes perdidos num 1-p.o.c., no sistema $M/M/1/10$ com taxa de serviço unitária, $P(S_1 \leq 3, L_1 \leq 3)$, em função da taxa de utilização.

Como esperado, para qualquer uma das políticas de bloqueio, à medida que a taxa de utilização aumenta, diminuiu o valor da função de distribuição conjunta no ponto considerado. Esta diminuição é mais

acentuada nas filas $M/M/1/10$ com bloqueio tradicional em que os clientes entram no sistema enquanto há lugares disponíveis. Em contraste, entre os sistemas $M/M/1/10$ que permitem desistências quando há ainda lugares disponíveis na fila, os sistemas com bloqueio reverso são os que apresentam maior probabilidade conjunta nesse ponto.

Agradecimentos

Este trabalho foi elaborado com o apoio parcial da Fundação para a Ciência e a Tecnologia (FCT) pelos projetos UID/Multi/04621/2013 e UID/Multi/04621/2019.

Referências

- [1] Jain, N.K., Kumar, R., Som, B. K. (2014). An $M/M/1/N$ Queuing System with Reverse Balking. *American Journal of Operational Research* 4(2), 17–20.
- [2] Kumar, R., Sharma, S. (2018). Transient Analysis of an $M/M/c$ Queuing System with Balking and Retention of Reneging Customers. *Communications in Statistics – Theory and Methods* 47(6), 1318–1327.
- [3] Kulkarni, V.G. (1995). *Modeling and Analysis of Stochastic Systems*. Chapman and Hall, Londres.
- [4] Haight, F. A. (1957). Queuing with balking I. *Biometrika* 44(3-4), p. 360–369.
- [5] Ancker, C. J., Gafarian, A. V. (1963). Some queuing problems with balking and reneging II. *Operations Research* 11(6), p. 928–937.
- [6] Guha, D., Goswami, V., Banik, A. D. (2016). Algorithmic computation of steady-state probabilities in an almost observable $GI/M/c$ queue with or without vacations under state dependent balking and reneging. *Applied Mathematical Modelling*, 40, 4199–4219.

Generalização do estimador de Hill, baseada na média de Lehmer: Resultados adicionais

Ivanilda Cabral

CMA, Universidade Nova de Lisboa

Universidade de Cabo Verde, *ivanilda.cabral@docente.unicv.edu.cv*

Frederico Caeiro

CMA e FCT, Universidade Nova de Lisboa, *fac@fct.unl.pt*

M. Ivette Gomes

CEAUL e DEIO, Universidade de Lisboa, *ivette.gomes@fc.ul.pt*

Palavras-chave: Índice de valores extremos; Estimação semi-paramétrica; Redução do viés.

Resumo: Este trabalho considera a generalização do estimador de Hill, baseado na média de Lehmer, introduzido em [2, 3] e posteriormente estudado em [20]. Removemos a componente dominante de viés assintótico deste estimador e apresentamos algumas das suas propriedades assintóticas sob a validade duma condição de variação regular de terceira ordem. As propriedades para amostras de pequena dimensão são obtidas através do método de simulação de Monte-Carlo.

1 Introdução

Seja $\{X_n\}_{n \in \mathbb{N}}$ uma sucessão de variáveis aleatórias (v.a.'s) independentes e identicamente distribuídas (*i.i.d.*) de um modelo F . Admita-se que F pertence ao domínio de atração para máximos de uma função de distribuição G , tal que, $G_\xi(x) = \exp\{-(1+\xi x)_+^{-1/\xi}\}$, $\xi \in \mathbb{R}$, $x_+ = \max(x, 0)$, ou seja, que $F \in D(G_\xi)$ com $\xi > 0$. A função

G é a distribuição de valores extremos e o parâmetro de forma, ξ , é usualmente conhecido por índice de valores extremos (EVI, do inglês *extreme value index*), sendo o parâmetro que se pretende estimar. Este parâmetro está diretamente ligado ao peso da cauda do modelo F , isto é, quanto maior for o valor de ξ , mais pesada a cauda $1 - F$. Esta distribuição G representa de modo unificado as três possíveis distribuições limite max-estáveis: Weibull ($\xi < 0$), Gumbel ($\xi = 0$) ou Fréchet ($\xi > 0$). Se $\xi > 0$,

$$F \in D(G_\xi) \iff 1 - F \in RV_{-1/\xi} \iff U \in RV_\xi, \quad (1)$$

onde U representa a função quantil definida por $U(t) = F^\leftarrow(1 - 1/t)$, $t \geq 1$, com $F^\leftarrow(t) = \inf\{x : F(x) \geq t\}$ a inversa generalizada de F e RV_α a classe das funções de variação regular em infinito de índice α , isto é, a classe das funções mensuráveis positivas $f(\cdot)$ tais que $f(tx)/f(t) \xrightarrow[t \rightarrow \infty]{} x^\alpha$, $\forall x > 0$, ([10]).

Na secção 2 apresentamos alguns estimadores clássicos e de viés reduzido do índice de valores extremos positivo. As distribuições assintóticas dos mesmos são obtidas com uma condição de segunda e de terceira ordem. Seguidamente, na mesma secção, fazemos a comparação, em níveis ótimos, entre os estimadores de viés reduzido. A secção 3 destina-se a um estudo de simulação de Monte Carlo para obter o comportamento dos estimadores, em estudo, para amostras dos modelos Fréchet e Burr.

2 Estimação do índice de valores extremos

Nos modelos de cauda pesada, $\xi > 0$, o estimador de Hill [18],

$$\hat{\xi}^H(k) = \frac{1}{k} \sum_{i=1}^k (\ln X_{n-i+1:n} - \ln X_{n-k:n}), \quad k = 1, 2, \dots, n-1, \quad (2)$$

é um dos primeiros e mais conhecidos estimador do índice de variação regular. Este estimador é consistente caso $F \in D(G_\xi)$ com $\xi > 0$

e k represente uma sequência intermédia, isto é, uma sequência de valores inteiros ($1 \leq k \leq n - 1$) verificando

$$k = k_n \longrightarrow \infty \quad \text{e} \quad k/n \longrightarrow 0, \quad n \rightarrow \infty. \quad (3)$$

A condição relativa a velocidade de convergência de $U(tx)/U(t)$ para x^ξ ,

$$\lim_{t \rightarrow \infty} \frac{\ln U(tx) - \ln U(t) - \xi \ln x}{A(t)} = \frac{x^\rho - 1}{\rho}, \quad \forall x > 0 \quad (4)$$

onde $\rho \leq 0$ é um parâmetro de forma de segunda ordem que mede a velocidade de convergência, é necessária para obter a distribuição limite do estimador $\hat{\xi}^H(k)$. Neste trabalho, considera-se que F pertence à vasta classe de modelos de Hall [17] para a qual é válida a parametrização $A(t) = \xi \beta t^\rho$ com $\beta \neq 0$ parâmetro de “escala” de segunda ordem e $\rho < 0$.

Assumindo as condições (3) e (4), o estimador de Hill possui a seguinte representação assintótica em distribuição

$$\sqrt{k}(\hat{\xi}^H(k) - \xi) \stackrel{d}{=} \xi Z_k + \frac{\sqrt{k}A(n/k)}{1 - \rho}(1 + o_p(1)), \quad (5)$$

onde $Z_k = \sum_{i=1}^k (E_i - 1)/\sqrt{k}$ é assintoticamente normal padrão e $E_i, i = 1, 2, \dots, k$ é uma sucessão de v.a.’s exponenciais unitárias independentes. Este estimador exhibe usualmente um viés assintótico acentuado quando k , o número de estatísticas ordinais de topo envolvidas na estimação, aumenta. A dificuldade na escolha do nível ótimo, nível que minimize o erro quadrático médio, levou muitos autores a considerarem outros estimadores do EVI. Essa dificuldade foi ultrapassada com a introdução de estimadores MVRB (do inglês “minimum variance reduced bias”). Nestes estimadores o termo dominante do viés do estimador de Hill, $A(n/k)/(1 - \rho) = \xi \beta (n/k)^\rho/(1 - \rho)$, é estimado e removido de modo adequado sem alterar o valor da sua variância assintótica. O estimador Corrected Hill, denotado CH, é o mais simples estimador MVRB

e foi introduzido em [5]. Este estimador é dado por:

$$\hat{\xi}_{\hat{\beta}, \hat{\rho}}^{CH}(k) = \hat{\xi}^H(k) \left(1 - \frac{\hat{\beta}}{1 - \hat{\rho}} \left(\frac{n}{k} \right)^{\hat{\rho}} \right), \quad k = 1, 2, \dots, n-1. \quad (6)$$

Apesar do referido estimador CH resultar da redução do termo dominante do viés assintótico do estimador de Hill, este estimador é geralmente enviesado para valores de k mais elevados (ver, por exemplo, a aplicação a dados reais em [15]). O viés é usualmente positivo. Penalva et al., [20] estudaram a seguinte generalização do estimador de Hill, baseada na média de Lehmer,

$$\hat{\xi}^{L_\alpha}(k) \equiv L_\alpha(k) := \frac{M_n^{(\alpha)}(k)}{\alpha M_n^{(\alpha-1)}(k)}, \quad \alpha > 0.5, \quad (7)$$

onde

$$M_n^{(\alpha)}(k) = \frac{1}{k} \sum_{i=1}^k (\ln X_{n-i+1:n} - \ln X_{n-k:n}), \quad k = 1, 2, \dots, n-1. \quad (8)$$

A classe de estimadores em (7) coincide com o estimador de Hill para $\alpha = 1$. Nesta generalização, o termo dominante do viés é de ordem $A(n/k)$.

Como o estimador baseado na média de Lehmer, $L_\alpha(k)$, é enviesado para qualquer α , Penalva et al., [20] consideraram a redução direta da componente dominante do viés através dos seguintes estimadores de viés reduzido (RB) de Lehmer,

$$\hat{\xi}_\alpha^{L^{RB}}(k) = L_\alpha^{RB}(k) := L_\alpha(k) \left(1 - \frac{\hat{\beta}(n/k)^{\hat{\rho}}}{(1 - \hat{\rho})^\alpha} \right), \quad \alpha > 0.5. \quad (9)$$

Escolhendo $\alpha = 1$, obtemos o estimador Corrected Hill em (6).

Os estimadores habitualmente usados para estimar ρ e β são aqueles que foram introduzidos em [8] e [12], respetivamente. Podemos encontrar algoritmos para a estimação de (ρ, β) em [13], entre outros.

Primeiramente, vamos estudar o estimador de Lehmer, introduzido em (9) e neste estudo vamos considerar os parâmetros de segunda ordem desconhecidos e conhecidos, isto é,

$$\hat{\xi}_{\alpha}^{\text{L}^{\text{RB}*}}(k) = \text{L}_{\alpha}^{\text{RB}*}(k) := \text{L}_{\alpha}(k) \left(1 - \frac{\beta(n/k)^{\rho}}{(1-\rho)^{\alpha}} \right), \quad \alpha > 0.5 \quad (10)$$

No que se segue, considera-se as seguintes notações:

$$\hat{\xi}^{\text{CH}}(k) = \hat{\xi}_{\beta,\rho}^{\text{CH}}(k), \quad \text{e} \quad \hat{\xi}_{\alpha}^{\text{L}^{\text{RB}}}(k) = \hat{\xi}_{\beta,\rho}^{\text{L}^{\text{RB}}}_{\alpha}(k). \quad (11)$$

2.1 Propriedades assintóticas

Vamos, nesta secção, obter as propriedades assintóticas dos estimadores H, CH, L_{α} , $\text{L}_{\alpha}^{\text{RB}}$ e $\text{L}_{\alpha}^{\text{RB}*}$ em (2), (6), (7), (9) e (10), respetivamente. Sob a validade da condição de segunda ordem em (4) e de k intermédio em (3), tal que $\lambda_A \xrightarrow[n \rightarrow +\infty]{d} \sqrt{k}A(n/k)$ finito, garantimos a normalidade assintótica dos referidos estimadores H, CH, L_{α} , $\text{L}_{\alpha}^{\text{RB}}$ e $\text{L}_{\alpha}^{\text{RB}*}$. Ou seja,

$$\lim_{n \rightarrow +\infty} \sqrt{k} \left(\hat{\xi}^{\bullet}(k) - \xi \right) \stackrel{d}{=} N(\lambda_A b_{\bullet}, \sigma_{\bullet}^2) \quad (12)$$

onde $N(\mu, \sigma^2)$ representa a normal padrão de média μ e variância σ^2 . Portanto, o viés e a variância dos estimadores clássicos são os seguintes:

$$b_{\text{H}} = \frac{1}{1-\rho}, \quad b_{\text{L}_{\alpha}} = \frac{1}{(1-\rho)^{\alpha}}, \quad b_{\text{L}_{\alpha}^{\text{RB}*}} = 0, \quad (13)$$

$$\sigma_{\text{H}}^2 = \xi^2, \quad \sigma_{\text{L}_{\alpha}}^2 = \frac{\xi^2 \Gamma(2\alpha-1)}{\Gamma^2(\alpha)} = \sigma_{\text{L}_{\alpha}^{\text{RB}*}}^2 \quad (14)$$

O termo dominante de viés dos estimadores de viés reduzido, CH e $\text{L}_{\alpha}^{\text{RB}}$ é nulo, ou seja, $b_{\text{CH}} = b_{\text{L}_{\alpha}^{\text{RB}}} = 0$, se os parâmetros ρ e β forem consistentemente estimados por $\hat{\rho}$ e $\hat{\beta}$, respetivamente, com $\hat{\rho} - \rho = o_p(1/\ln n)$. Já a variância dos referidos estimadores de viés

reduzido é igual às variâncias dos seus respetivos estimadores clássicos, isto é, $\sigma_{CH}^2 = \sigma_H^2$ e $\sigma_{L_{\alpha}^{RB}}^2 = \sigma_{L_{\alpha}}^2$.

Para obter mais informação sobre o viés do estimador em estudo, vamos impor a validade da seguinte condição de terceira ordem: existe uma função $B(t)$ que mede a velocidade de convergência de (4) tal que, para todo o $x > 0$,

$$\lim_{t \rightarrow \infty} \frac{\frac{\ln U(tx) - \ln U(t) - \xi \ln x}{A(t)} - \frac{x^\rho - 1}{\rho}}{B(t)} = \frac{x^{2\rho} - 1}{2\rho}, \quad (15)$$

onde $\beta' \leq 0$ é um parâmetro de terceira ordem e $|B(t)| \in RV_\rho$. Vamos ainda assumir que $A(t) = \xi \beta t^\rho$, $B(t) = \beta' t^\rho$, com $\beta, \beta' \neq 0$, e $\rho < 0$. Consideremos a notação $\zeta = \beta'/\beta$.

A distribuição assintótica do estimador $\hat{\xi}_{\alpha}^{RB*}(k)$, definido em (11), é enunciada num contexto de terceira ordem na seguinte proposição.

Proposição 2.1 *Consideremos que são válidas as condições (3) e (15) e que conhecemos os valores dos parâmetros de segunda ordem ρ e β . Então,*

$$\sqrt{k}(\hat{\xi}^{\bullet}(k) - \xi) \xrightarrow[n \rightarrow +\infty]{d} N(0, \sigma_{\bullet}^2) + b_{\bullet} \sqrt{k} A(n/k) + c_{\bullet} \sqrt{k} A^2(n/k)(1 + o_p(1))$$

onde podemos designar \bullet como H , L_{α} ou L_{α}^{RB*} , b_{\bullet} e σ_{\bullet}^2 estão representados em (12) e (14) e, c_{\bullet} é dado por:

$$c_H = \frac{1}{\xi(1-2\rho)}, \quad c_{L_{\alpha}} = \frac{1}{\xi\rho} \left(\frac{\zeta\rho+1-2\rho}{(1-2\rho)^{\alpha}} - \frac{1-\rho(1-\rho)^{\alpha-1}}{(1-2\rho)^{2\alpha-1}} \right), \quad (16)$$

$$c_{L_{\alpha}^{RB*}} = \frac{1}{\xi\rho} \left(\frac{\zeta\rho+1-2\rho}{(1-2\rho)^{\alpha}} - \frac{1-\rho(1-\rho)^{\alpha}}{(1-2\rho)^{2\alpha}} \right). \quad (17)$$

Proposição 2.2 *Nas condições da Proposição 2.1, consideremos estimadores consistentes $(\hat{\beta}, \hat{\rho})$ dos parâmetros (β, ρ) ambos calculados num nível k_1 tal que $k = o(k_1)$ e, assumimos $(\hat{\rho} - \rho) \ln(n) =$*

$o_p(1)$. Se $(\hat{\beta} - \beta)/\beta \stackrel{p}{\sim} -(\hat{\rho} - \rho) \ln(n/k_1)$, então

$$\hat{\xi}^{L^{RB}}_{\alpha}(k) - \hat{\xi}^{L^{RB+}}_{\alpha}(k) \stackrel{p}{\sim} -\frac{A(n/k)}{(1-\rho)^{\alpha}}(\hat{\rho} - \rho) \ln(k_1/k).$$

Consequentemente, $\hat{\xi}^{L^{RB}}_{\alpha}(k)$ é consistente se $(\hat{\rho} - \rho) \ln(k/k_1) = o_p(1/A(n/k))$ e possui a distribuição normal assintótica se $(\hat{\rho} - \rho) \ln(k/k_1) = o_p(1/\sqrt{k}A(n/k))$

Dem.: Através da parameterização $A(n/k) = \xi\beta(n/k)^{\rho}$ e da expansão em série de Taylor,

$$\begin{aligned} & \frac{\hat{\beta}(n/k)^{\hat{\rho}}}{(1-\hat{\rho})^{\alpha}} \\ &= \frac{\beta(n/k)^{\rho}}{(1-\rho)^{\alpha}} + \frac{A(n/k)}{\xi(1-\rho)^{\alpha}}(\hat{\beta}(k_1) - \beta)(1 + o_p(1)) \\ &+ (n/k)^{\rho} \ln(n/k) \frac{\beta}{(1-\rho)^{\alpha}}(\hat{\rho}(k_1) - \rho)(1 + o_p(1)) \end{aligned}$$

e usando a aproximação $\frac{\hat{\xi}^{L_{\alpha}}(k)}{\xi} = 1$, obtemos o resultado pretendido. Ou seja,

$$\begin{aligned} \hat{\xi}^{L^{RB}}_{\alpha}(k) &= \hat{\xi}^{L_{\alpha}}(k) \left(1 - \frac{\hat{\beta}(n/k)^{\hat{\rho}}}{(1-\hat{\rho})^{\alpha}} \right) \\ &\stackrel{d}{=} \hat{\xi}^{L^{RB*}}_{\alpha}(k) - \frac{A(n/k)}{(1-\rho)^{\alpha}}(\hat{\rho} - \rho) \ln(k_1/k) \end{aligned}$$

■

2.2 Comparação assintótica em níveis ótimos

Nesta subsecção, vamos comparar assintoticamente o estimador L^{RB}_{α} com o estimador CH, nos respetivos níveis ótimos. A comparação será feita de modo similar à comparação em [7] para estimadores

clássicos e em [6] para estimadores de viés reduzido. Aqui consideramos modelos de cauda pesada que verificam (15), com $\rho = \rho' < 0$, $A(t) = \xi \beta t^\rho$ e $B(t) = \beta' t^\rho = \zeta A(t)/\xi$, com $\zeta = \beta'/\beta$. Estas condições são verificadas por vários modelos de cauda pesada ([6]). Nesta subsecção vamos usar os seguintes modelos:

- Fréchet, com função de distribuição $F(x) = \exp(-x^{-1/\xi})$, $x \geq 0$, $\xi > 0$ ($\rho' = \rho = -1$, $\beta = 0.5$ e $\beta' = 5/6$);
- Burr com função de distribuição $F(x) = 1 - (1 + x^{-\rho/\xi})^{1/\rho}$, $x \geq 0$, $\xi > 0$, $\rho < 0$ ($\rho' = \rho < 0$ e $\beta = \beta' = 1$);

Vamos denotar por $\hat{\xi}^\bullet(k)$ um qualquer dos dois estimadores MVRB em (11). Então temos

$$\hat{\xi}^\bullet(k) \stackrel{d}{=} \xi + \frac{\sigma_\bullet}{\sqrt{k}} Z_k + b_\bullet A^2(n/k)(1 + o_p(1)), \quad (18)$$

onde Z_k é a sucessão de variáveis aleatórias introduzida em (5). A variância e o viés assintótico do estimador $\hat{\xi}^\bullet(k)$ são dados por σ_\bullet^2/k e $b_\bullet A^2(n/k)$, respetivamente. O erro quadrático médio assintótico (AMSE) é então dado por $\text{AMSE}[\hat{\xi}^\bullet(k)] = \sigma_\bullet^2/k + b_\bullet^2 A^4(n/k)$. Considerando dois estimadores $\hat{\xi}^{(1)}(k)$ e $\hat{\xi}^{(2)}(k)$ para os quais é válida a representação em (18), calculados nos respetivos níveis óptimos, $k_0^{(j)} := \arg \min_k \text{AMSE}[\hat{\xi}^{(j)}(k)]$, e a notação $\hat{\xi}_0^{(j)} = \hat{\xi}^{(j)}(k_0^{(j)})$, $j = 1, 2$. A eficiência relativa assintótica, ARE (do inglês “Asymptotic Relative Efficiency”), de $\hat{\xi}_0^{(1)}$ relativamente a $\hat{\xi}_0^{(2)}$ é obtida através do seguinte indicador ([6]):

$$ARE_{1|2} = ARE_{\hat{\xi}_0^{(1)}|\hat{\xi}_0^{(2)}} = \left[\left(\frac{\sigma_2}{\sigma_1} \right)^{-4\rho} \left| \frac{b_2}{b_1} \right| \right]^{\frac{1}{1-4\rho}}.$$

Quanto maior for o valor de $ARE_{1|2}$, melhor é o estimador $\hat{\xi}_0^{(1)}$. Para os estimadores em estudo, $\sigma_{CH}^2 = \xi^2$, $\sigma_{L_\alpha^{RB}}^2 = \xi^2 \frac{\Gamma(2\alpha-1)}{\Gamma^2(\alpha)}$, $b_{CH} =$

$\frac{\zeta}{\xi(1-2\rho)} - \frac{1}{\xi(1-\rho)^2}$ e $b_{L_{\alpha}^{RB}} = \frac{1}{\rho\xi} \left(\frac{1-2\rho+\tau\rho}{(1-2\rho)^{\alpha}} + \frac{\rho(1-\rho)^{\alpha}-1}{(1-\rho)^{2\alpha}} \right)$. Consequentemente,

$$ARE_{L_{\alpha}^{RB}|CH} = \left\{ \left(\frac{\Gamma^2(\alpha)}{\Gamma(2\alpha-1)} \right)^{-2\rho} \left| \frac{\zeta\rho(1-\rho)^{2\alpha}(1-2\rho)^{\alpha-1} - \rho(1-\rho)^{2\alpha-2}(1-2\rho)^{\alpha}}{(\zeta\rho+1-2\rho)(1-\rho)^{2\alpha} - (1-\rho(1-\rho)^{\alpha})(1-2\rho)^{\alpha}} \right| \right\}^{\frac{1}{1-4\rho}}.$$

Fizemos a escolha do α a partir de observações do gráfico dos referidos modelos, Fréchet e Burr, que apresentaram maior estabilidade junto ao valor do $\xi = 0.5$. Neste sentido, os valores de $\alpha = 1.1$ e $\alpha = 1.2$ foram os escolhidos como referência no cálculo da eficiência assintótica.

Para o modelo Fréchet, os valores da eficiência relativa assintótica do estimador de Lehmer relativamente ao estimador CH são de 0.99 para $\alpha = 1.1$ e de 1.1 para $\alpha = 1.2$. Isso significa que, para este modelo, o estimador $\hat{\xi}_{L_{\alpha}^{RB}}^{\text{L}^{\text{RB}}}(k)$ é assintoticamente mais eficiente do que o estimador $\hat{\xi}^{CH}(k)$, no respetivo nível ótimo, para $\alpha = 1.2$.

Relativamente ao modelo Burr, com $\rho = -0.75$ e $\rho = -1$, o valor do indicador $ARE_{L_{\alpha}^{RB}|CH}$ é sempre superior a 1 tanto para $\alpha = 1.1$ como para $\alpha = 1.2$. Portanto, para este modelo, o estimador $\hat{\xi}_{L_{\alpha}^{RB}}^{\text{L}^{\text{RB}}}(k)$ é assintoticamente mais eficiente do que o estimador $\hat{\xi}^{CH}(k)$, no respetivo nível ótimo.

3 Estudo de Simulação

3.1 Metodologia e resultados

Apresenta-se, nesta secção, um estudo de simulação para analisar o comportamento de $\hat{\xi}^H$, $\hat{\xi}^{CH}$ e $\hat{\xi}_{L_{\alpha}^{RB}}^{\text{L}^{\text{RB}}}$ apresentados em (2), (6) e (9) respetivamente. Nesse estudo, foram geradas 1000 amostras de dimensão n , para diferentes valores de n , dos modelos Fréchet com $\xi = 0.5$ e Burr com $\xi = 0.5$ e $\rho = -1$ e -0.75 . Para cada amostra de dimensão n , foram calculadas as estimativas $\hat{\xi}_i^{\bullet}(k)$, $k = 1, 2, \dots, n-1$, $i = 1, 2, \dots, 1000$, que permitiram obter estimativas do valor médio

(E) e da raíz quadrada do erro quadrático médio (RMSE) dados por

$$E[\hat{\xi}^\bullet(k)] = \sum_{i=1}^{1000} \frac{\hat{\xi}_i^\bullet(k)}{1000} \quad \text{e} \quad RMSE[\hat{\xi}^\bullet(k)] = \sqrt{\sum_{i=1}^{1000} \frac{(\hat{\xi}_i^\bullet(k) - \xi)^2}{1000}}. \quad (19)$$

As Figuras 1, 2 e 3 representam os valores simulados de $E[\hat{\xi}^\bullet(k)]$ e $RMSE[\hat{\xi}^\bullet(k)]$ em (19), para amostras de dimensão $n = 1000$ dos modelos em estudo. Com base nos valores dados por (19), determinámos $\hat{k}_0 = \arg \min_k RMSE[\hat{\xi}^\bullet(k)]$ com o qual obtivemos

$$E[\hat{\xi}_0] = E[\hat{\xi}^\bullet(\hat{k}_0)] \quad \text{e} \quad RMSE[\hat{\xi}_0] = RMSE[\hat{\xi}^\bullet(\hat{k}_0)]. \quad (20)$$

Na Tabela 1 foram apresentados os valores dos indicadores dados em (20), para várias dimensões de amostras e os modelos e valores de parâmetros usados nas figuras.

3.2 Conclusões

Analizando os gráficos, em todas as dimensões de amostras consideradas, nos modelos Fréchet e Burr, constata-se que há uma ligeira melhoria dos valores médios de L_α^{RB} comparativamente com os valores médios de CH para $\alpha = 1.1$ e $\alpha = 1.2$, no nível ótimo. Para esses valores de α , o estimador de Lehmer tende a ser mais estável em torno do $\xi = 0.5$ considerado.

Quanto às tabelas, observa-se que para os modelos em estudo, nas dimensões de amostras consideradas, os valores médios simulados dos estimadores CH e L_α^{RB} , no nível ótimo, estão mais próximas do verdadeiro valor de ξ do que os valores médios simulados do estimador de Hill no seu nível ótimo. No entanto, para os modelos Fréchet com $\xi = 0.5$ e Burr com $(\xi, \rho) = (0.5, -1)$, em todas as dimensões de amostras consideradas, os valores médios simulados do estimador L_α^{RB} , no nível ótimo, são os que se encontram mais próximos do verdadeiro valor de ξ . O mesmo não acontece para o modelo Burr com $(\xi, \rho) = (0.5, -0.75)$ onde o estimador CH apresenta o melhor valor médio simulado.

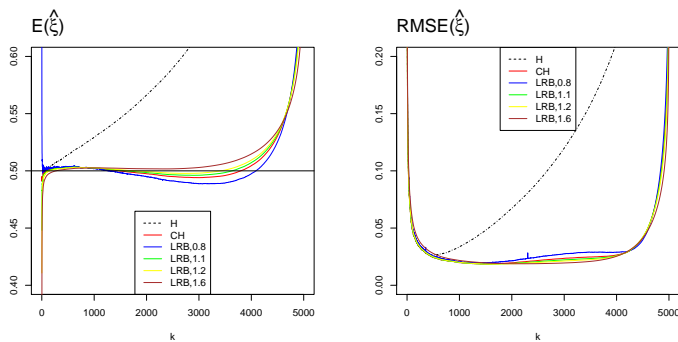


Figura 1: Valores médios (esquerda) e RMSE (direita) simulados para amostras de dimensão $n = 5000$ do modelo *Fréchet* com $\xi = 0.5$.

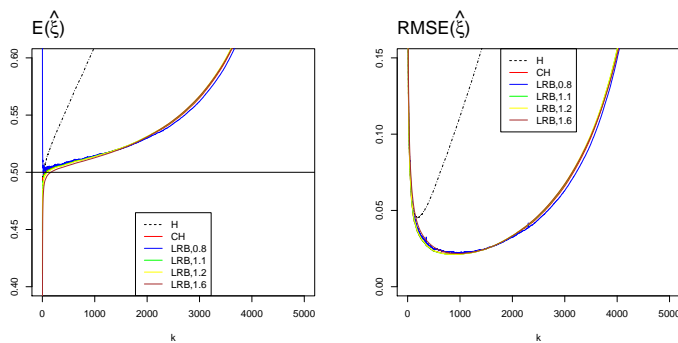


Figura 2: Valores médios (esquerda) e RMSE (direita) simulados para amostras de dimensão $n = 5000$ do modelo *Burr* com $(\xi, \rho) = (0.5, -0.75)$.

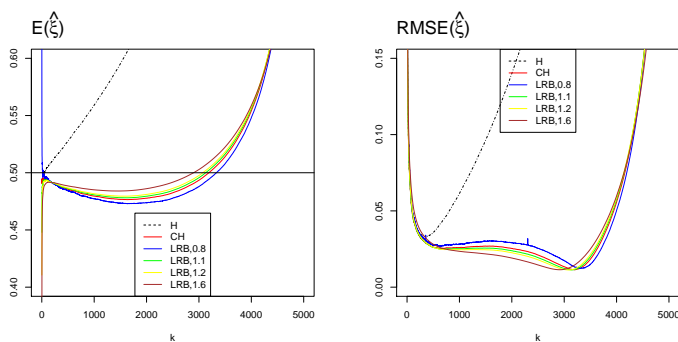


Figura 3: Valores médios (esquerda) e RMSE (direita) simulados para amostras de dimensão $n = 5000$ do modelo *Burr* com $(\xi, \rho) = (0.5, -1)$.

Tabela 1: Estimativas do valor esperado, no nível ótimo, dos estimadores do índice de valores extremos, $\hat{\xi}^H(k)$, $\hat{\xi}^{CH}(k)$ e $\hat{\xi}^{L_{\alpha}^{RB}}(k)$.

Fréchet $\xi = 0.5$ ($\rho = -1$, $\beta = 0.5$)							
n	100	200	500	1000	2000	5000	10000
$\hat{\xi}_0^H$	0.5601	0.5377	0.5314	0.5268	0.5209	0.5139	0.5120
$\hat{\xi}_0^{CH}$	0.4907	0.4955	0.4962	0.4999	0.5003	0.5005	0.4991
$\hat{\xi}_{0,(0.8)}^{L_{\alpha}^{RB}}$	0.4837	0.4855	0.4947	0.4995	0.4994	0.4992	0.4976
$\hat{\xi}_{0,(1.1)}^{L_{\alpha}^{RB}}$	0.4957	0.4981	0.4963	0.5002	0.5007	0.5004	0.4981
$\hat{\xi}_{0,(1.2)}^{L_{\alpha}^{RB}}$	0.4974	0.4999	0.4974	0.5001	0.5010	0.5005	0.4991
$\hat{\xi}_{0,(1.6)}^{L_{\alpha}^{RB}}$	0.5075	0.5064	0.5049	0.5025	0.5019	0.5017	0.5007
Burr $\xi = 0.5$, $\rho = -0.75$ ($\beta = 0.5$)							
$\hat{\xi}_0^H$	0.6052	0.5765	0.5467	0.5427	0.5368	0.5259	0.5220
$\hat{\xi}_0^{CH}$	0.5383	0.5309	0.5202	0.5204	0.5153	0.5139	0.5100
$\hat{\xi}_{0,(0.8)}^{L_{\alpha}^{RB}}$	0.5409	0.5324	0.5207	0.5199	0.5155	0.5141	0.5100
$\hat{\xi}_{0,(1.1)}^{L_{\alpha}^{RB}}$	0.5368	0.5302	0.5200	0.5202	0.5151	0.5114	0.5100
$\hat{\xi}_{0,(1.2)}^{L_{\alpha}^{RB}}$	0.5356	0.5295	0.5211	0.5201	0.5154	0.5111	0.5090
$\hat{\xi}_{0,(1.6)}^{L_{\alpha}^{RB}}$	0.5396	0.5318	0.5224	0.5197	0.5150	0.5125	0.5090
Burr $\xi = 0.5$, $\rho = -1$ ($\beta = 0.5$)							
$\hat{\xi}_0^H$	0.5702	0.5427	0.5415	0.5273	0.5250	0.5189	0.5130
$\hat{\xi}_0^{CH}$	0.5060	0.4997	0.4860	0.4806	0.4982	0.5000	0.5000
$\hat{\xi}_{0,(0.8)}^{L_{\alpha}^{RB}}$	0.5043	0.4982	0.4839	0.4785	0.4952	0.4991	0.5000
$\hat{\xi}_{0,(1.1)}^{L_{\alpha}^{RB}}$	0.5064	0.5004	0.4904	0.4859	0.4992	0.5000	0.5000
$\hat{\xi}_{0,(1.2)}^{L_{\alpha}^{RB}}$	0.5067	0.5035	0.4928	0.4883	0.4971	0.5000	0.5000
$\hat{\xi}_{0,(1.6)}^{L_{\alpha}^{RB}}$	0.5108	0.5095	0.5042	0.4978	0.4991	0.5000	0.5000

Tabela 2: Estimativas da raiz quadrada do erro quadrático médio, no nível ótimo, dos estimadores do índice de valores extremos, $\hat{\xi}^H(k)$, $\hat{\xi}^{CH}(k)$ e $\hat{\xi}^{L\text{RB}}_{\alpha}(k)$.

Fréchet $\xi = 0.5$ ($\rho = -1$, $\beta = 0.5$)							
n	100	200	500	1000	2000	5000	10000
$\hat{\xi}_0^H$	0.1040	0.0812	0.0595	0.0464	0.0368	0.0263	0.0210
$\hat{\xi}_0^{CH}$	0.0888	0.0692	0.0526	0.0388	0.0285	0.0186	0.0132
$\hat{\xi}_{0,(0.8)}^{L\text{RB}}_{\alpha}$	0.0942	0.0732	0.0544	0.0400	0.0296	0.0191	0.0138
$\hat{\xi}_{0,(1.1)}^{L\text{RB}}_{\alpha}$	0.0872	0.0678	0.0524	0.0388	0.0283	0.0185	0.0131
$\hat{\xi}_{0,(1.2)}^{L\text{RB}}_{\alpha}$	0.0861	0.0668	0.0520	0.0389	0.0284	0.0185	0.0131
$\hat{\xi}_{0,(1.6)}^{L\text{RB}}_{\alpha}$	0.0854	0.0661	0.0511	0.0389	0.0290	0.0188	0.0131
Burr $\xi = 0.5$, $\rho = -0.75$ ($\beta = 0.5$)							
$\hat{\xi}_0^H$	0.1671	0.1292	0.0937	0.0745	0.0604	0.0451	0.0360
$\hat{\xi}_0^{CH}$	0.1021	0.0748	0.0515	0.0380	0.0281	0.0213	0.0160
$\hat{\xi}_{0,(0.8)}^{L\text{RB}}_{\alpha}$	0.1092	0.0796	0.0537	0.0397	0.0291	0.0220	0.0170
$\hat{\xi}_{0,(1.1)}^{L\text{RB}}_{\alpha}$	0.1005	0.0737	0.0513	0.0378	0.0281	0.0212	0.0160
$\hat{\xi}_{0,(1.2)}^{L\text{RB}}_{\alpha}$	0.0910	0.0732	0.0512	0.0378	0.0282	0.0212	0.0160
$\hat{\xi}_{0,(1.6)}^{L\text{RB}}_{\alpha}$	0.1001	0.0733	0.0525	0.0389	0.0294	0.0217	0.0170
Burr $\xi = 0.5$, $\rho = -1$ ($\beta = 0.5$)							
$\hat{\xi}_0^H$	0.1312	0.1032	0.0743	0.0573	0.0440	0.0330	0.0261
$\hat{\xi}_0^{CH}$	0.0900	0.0725	0.0550	0.0429	0.0280	0.0110	0.0070
$\hat{\xi}_{0,(0.8)}^{L\text{RB}}_{\alpha}$	0.0955	0.0769	0.0584	0.0453	0.0310	0.0120	0.0070
$\hat{\xi}_{0,(1.1)}^{L\text{RB}}_{\alpha}$	0.0886	0.0711	0.0539	0.0418	0.0270	0.0110	0.0070
$\hat{\xi}_{0,(1.2)}^{L\text{RB}}_{\alpha}$	0.0877	0.0701	0.0527	0.0415	0.0260	0.0110	0.0070
$\hat{\xi}_{0,(1.6)}^{L\text{RB}}_{\alpha}$	0.0872	0.0684	0.0502	0.0373	0.0240	0.0110	0.0080

Relativamente ao RMSE, no nível ótimo, o estimador L_{α}^{RB} apresenta o menor valor, comparativamente com o RMSE dos estimadores de Hill e CH nos modelos Fréchet com $\xi = 0.5$ e Burr com $(\xi, \rho) = (0.5, -1)$. A exceção é o modelo Burr com $(\xi, \rho) = (0.5, -0.75)$. Para detalhes adicionais sobre este tipo de estimadores de Lehmer, veja-se [19].

Agradecimentos

À FCT-UNL pelo apoio financeiro concedido a Ivanilda Cabral. Investigação parcialmente suportada pela FCT-Fundação para a Ciência e a Tecnologia, projectos PEst-OE/MAT/UI006/2014 (CEA/UL) e UID/MAT/00297/2013 (CMA/UNL).

Referências

- [1] Beirlant, J., Caeiro, F., Gomes, M.I. (2012). An Overview and Open Research Topics in Statistics of Univariate Extremes. *Revstat* 10(1), 1–31.
- [2] Caeiro, F. and Gomes, M. I. A class of asymptotically unbiased semi-parametric estimators of the tail index. *Test* 11(2), 345–364, 2002.
- [3] Caeiro, F. and Gomes, M. I. Bias reduction in the estimation of parameters of rare events. *Theory of Stochastic Processes* 8(24), 67–76, 2002.
- [4] Caeiro, F. and Gomes, M. I. Comparison of asymptotically unbiased extreme value index estimators: a Monte Carlo simulation study. *AIP Conference Proceedings* 1618, 551–554, 2014.
- [5] Caeiro, F., Gomes, M.I., Pestana, D. (2005). Direct Reduction of Bias of the Classical Hill Estimator. *Revstat* 3(2), 113–136.
- [6] Caeiro, F., Gomes, M.I. (2011). Asymptotic comparison at optimal levels of reduced-bias extreme value index estimators. *Statistica Neerlandica* 65(4), 462–488.
- [7] de Haan, L., Peng, L. (1998). Comparison of tail index estimators. *Statistica Neerlandica* 52(1), 60–70.

- [8] Fraga Alves, M.I., Gomes, M.I., de Haan, L. (2003). A new class of semiparametric estimators of the second order parameter. *Portugaliae Mathematica* 60(2), 193–213.
- [9] Geluk, J., de Haan, L. (1987). *Regular Variation, Extensions and Tauberian Theorems*. Tech. Report CWI Tract 40, Centre for Mathematics and Computer Science, Amsterdam, Netherlands.
- [10] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d’une série aléatoire. *Ann. Math.* 44, 423–453.
- [11] Gomes, M. I., Caeiro, F. and Figueiredo, F. Bias Reduction of a Tail Index Estimator Through an External Estimation of the second-order parameter. *Statistics* 38(6), 497–510, 2004.
- [12] Gomes, M.I., Martins, M.J. (2002). Asymptotically unbiased estimators of the tail index based on external estimation of the second order parameter. *Extremes* 5, 5–31.
- [13] Gomes, M.I., Pestana, D. (2007). A sturdy reduced-bias extreme quantile (VaR) estimator. *Journal American Statistical Association* 102, 280–292.
- [14] Gomes, M.I., Pestana, D., Caeiro, F. (2009). A Note on the Asymptotic Variance at Optimal Levels of Bias-Corrected Hill Estimator. *Statistics Probability Letters* 79(3), 295–303.
- [15] Gomes, M.I., Henriques-Rodrigues, L., Fraga Alves, M.I., Manjunath, B.G. (2013). Adaptive PORT-MVRB estimation: an empirical comparison of two heuristic algorithms. *J. Statist. Comput. Simul.* 83(6), 1129–1144.
- [16] Gomes M.I., Penalva, H., Caeiro, F. and Neves M.M. (2016). Non-reduced versus reduced-bias estimators of the extreme value index-efficiency and robustness. In Colubi, A., Blanco A. and Gatu C. (eds.), *Proceedings of COMPSTAT 2016: 22th International Conference on Computational Statistics*, Oviedo, Spain, 279–290.
- [17] Hall, P. (1982). On some Simple Estimates of an Exponent of Regular Variation. *J. R. Statist. Soc.* 44(1), 37–42.
- [18] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* 3, 1163–1174.

- [19] Penalva, H. (2017). *Contributos Computacionais e Metodológicos na Estimação do Índice de Valores Extremos*. Tese de Doutoramento em Matemática e Estatística, Instituto Superior de Agronomia.
- [20] Penalva, H., Gomes, M.I., Caeiro, F. and Neves, M.: A couple of non reduced bias generalized means in extreme value theory: an asymptotic comparison. Accepted in *REVSTAT*.

Modelos de sobrevivência aplicados à análise de acontecimentos múltiplos

Ivo Sousa-Ferreira

Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira e Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal, *ivo.ferreira@staff.uma.pt*

Ana Maria Abreu

Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira e Centro de Investigação em Matemática e Aplicações, *abreu@staff.uma.pt*

Cristina Rocha

Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal e Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal, *cmrocha@fc.ul.pt*

Palavras-chave: Acontecimentos múltiplos; Análise de sobrevivência; Extensões do modelo de Cox; Modelos paramétricos; *Software R*.

Resumo: Na modelação de acontecimentos múltiplos, uma abordagem muito utilizada consiste em desenvolver extensões do modelo semiparamétrico de Cox. Contudo, quando se considera que o conhecimento da distribuição do tempo é importante para o estudo, a abordagem paramétrica revela-se mais adequada. Neste trabalho, são apresentados dois novos modelos paramétricos baseados na distribuição de Weibull (modelos WNE e WE), como alternativas a duas das extensões do modelo de Cox para acontecimentos recorrentes (modelos AG e PWP). De forma a ilustrar as diferenças entre estas duas abordagens, é considerado um exemplo de aplicação com

dados simulados. Os resultados obtidos com os modelos propostos permitem acrescentar mais duas opções ao leque de escolhas de modelos para acontecimentos recorrentes.

1 Introdução

Em estudos de longa duração é cada vez mais frequente observar a ocorrência de vários acontecimentos para o mesmo indivíduo. Situações deste tipo surgem, por exemplo, nas ciências da saúde, quando se pretende analisar o tempo até sucessivas recaídas de uma doença; e nas ciências económicas, onde interessa avaliar as razões que levam à observação de insolvências de instituições bancárias num determinado país. Assim, tem-se verificado um crescente empenho em desenvolver metodologia estatística capaz de dar resposta a estes cenários complexos.

Na literatura, existem várias abordagens para modelar o tempo até à ocorrência de acontecimentos múltiplos [5]. Aquela que mais tem sido aplicada, dada a sua versatilidade, consiste em adaptar o modelo semiparamétrico de Cox [2] a esta situação. De facto, nas últimas quatro décadas têm sido propostas várias extensões deste modelo que visam ter em conta diversos aspetos relevantes num estudo desta natureza [3, 16], nomeadamente: acontecimentos do mesmo tipo (designados por acontecimentos recorrentes) ou de tipos diferentes; acontecimentos instantâneos ou duradouros; acontecimentos com riscos de ocorrência distintos; acontecimentos com uma estrutura de ordenação; e acontecimentos com uma estrutura de dependência.

Uma outra abordagem, ainda pouco trabalhada no contexto dos acontecimentos múltiplos, consiste na modelação totalmente paramétrica dos tempos até à ocorrência dos acontecimentos. Note-se que, nos modelos que são extensões do modelo de Cox, a função de risco subjacente não é especificada, o que pode constituir uma limitação. Com efeito, em certas circunstâncias, a estimação desta função é extremamente importante, em particular na área da saúde,

pois permite estudar a evolução de uma doença ao longo do tempo. Deste modo, os modelos paramétricos têm a vantagem de permitir a estimação da função de risco subjacente.

O principal propósito deste trabalho é dar a conhecer dois tipos de abordagem para a modelação de acontecimentos recorrentes: semiparamétrica *versus* paramétrica. Para tal, começa-se por referir a abordagem semiparamétrica, onde serão consideradas duas extensões do modelo de Cox e, em seguida, a abordagem paramétrica, onde serão apresentados dois novos modelos baseados na distribuição de Weibull. Por último, será apresentado um exemplo de aplicação com dados simulados de forma a ilustrar as diferenças entre estas abordagens.

2 Metodologia

Os modelos de riscos proporcionais constituem uma classe de modelos de regressão bem conhecida em Análise de Sobrevivência. Este tipo de modelos é caracterizado pela proporcionalidade entre as funções de risco respeitantes a indivíduos com diferentes valores das covariáveis. De facto, considerando T uma variável aleatória (v.a.) contínua e $\mathbf{z} = (z_1, \dots, z_p)'$ um vetor de covariáveis associado a um determinado indivíduo, o modelo de riscos proporcionais formulado com base na função de risco é usualmente escrito na forma

$$h(t; \mathbf{z}) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{z}), \quad t \geq 0, \quad (1)$$

em que $\boldsymbol{\beta}' = (\beta_1, \dots, \beta_p)$ é o vetor de parâmetros de regressão que representa o efeito (desconhecido) das p covariáveis, e $h_0(t)$ denota uma função não negativa, designada por função de risco subjacente. Quando não se especifica uma forma particular para esta função, isto é, quando $h_0(t)$ é uma função arbitrária, obtém-se o modelo semiparamétrico proposto por Sir David Cox [2]. Por outro lado, quando se admite que o tempo de vida segue uma determinada distribuição, o modelo resultante é totalmente paramétrico.

Antes de proceder à formulação dos modelos que serão considerados neste trabalho, interessa introduzir alguma notação adicional. Seja n o número de indivíduos em estudo, em que para cada um pode ser observado um máximo de K acontecimentos, e $X_{ik} = \min\{T_{ik}, C_{ik}\}$ o tempo em observação do indivíduo i ($i = 1, \dots, n$) correspondente ao acontecimento k ($k = 1, \dots, K$), onde T_{ik} e C_{ik} representam o seu verdadeiro tempo e o seu tempo de censura, respetivamente. A variável indicatriz que caracteriza o estado do i -ésimo indivíduo em relação ao acontecimento k é definida por $\delta_{ik} = I(T_{ik} \leq C_{ik})$, a qual toma o valor um quando o acontecimento k é observado e zero caso contrário. Por fim, denote-se por $\mathbf{z}_{ik}(t) = (z_{ik1}(t), \dots, z_{ikp}(t))'$ o vetor de p covariáveis (possivelmente dependentes do tempo) associado ao i -ésimo indivíduo referente ao acontecimento k .

2.1 Extensões do modelo semiparamétrico de Cox

O modelo de Cox é apropriado para analisar o tempo decorrido desde um instante inicial, bem definido, até à observação de um único acontecimento de interesse. Porém, no contexto dos acontecimentos múltiplos, o facto de se poder registar mais do que um tempo para cada indivíduo inviabiliza a aplicação direta deste modelo. Por essa razão, têm sido propostas várias extensões do modelo de Cox para analisar acontecimentos múltiplos e, em particular, acontecimentos de um único tipo que se repetem ao longo do tempo [16]. Este trabalho focar-se-á nesta segunda situação, sendo que duas das extensões do modelo de Cox que mais têm sido aplicadas para analisar acontecimentos recorrentes foram propostas por Andersen e Gill (AG) [1] e Prentice, Williams e Peterson (PWP) [11].

No que concerne à formulação dos modelos AG e PWP, para o i -ésimo indivíduo em estudo, as respetivas funções de risco são dadas por

$$h(t; \mathbf{z}_{ik}(t)) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{z}_{ik}(t)), \quad t \geq 0, \quad (2)$$

e

$$h(t; \mathbf{z}_{ik}(t)) = h_{0k}(t) \exp(\boldsymbol{\beta}' \mathbf{z}_{ik}(t)), \quad t \geq 0, \quad (3)$$

onde $h_0(t) \geq 0$ representa a função de risco subjacente comum a todos os acontecimentos, $h_{0k}(t) \geq 0$ é a função de risco subjacente específica do acontecimento k e $\beta = (\beta_1, \dots, \beta_p)'$ é o vetor de parâmetros de regressão. Note-se que estas extensões são na verdade modelos semiparamétricos pois a forma da função de risco subjacente não é especificada.

Em termos gerais, o modelo AG (2) foi proposto para o caso em que os acontecimentos ocorrem de forma ordenada e apresentam igual risco de ocorrerem, pelo que se considera uma função de risco subjacente comum a todos os acontecimentos. Neste modelo, considera-se que todos os indivíduos em estudo contribuem para o conjunto de risco de qualquer acontecimento, seja qual for o número de acontecimentos observados para cada indivíduo. Deste modo, diz-se que o conjunto de risco é não restritivo.

O modelo PWP (3) também surgiu para analisar acontecimentos ordenados, mas pressupõe que o risco de ocorrência de um acontecimento é afetado pela ocorrência do acontecimento que o antecede. Consequentemente, é necessário estratificar os indivíduos segundo a ordem pela qual os acontecimentos ocorrem. Assim, se for possível observar k acontecimentos, existirão k estratos ordenados, sendo que a cada um deles estará associada a função de risco subjacente $h_{0k}(t)$, $k = 1, \dots, K$. Note-se que neste modelo tanto pode ser obtida uma estimativa global dos parâmetros de regressão $\beta = (\beta_1, \dots, \beta_p)'$ (alternativa que adotamos de modo a poder ser comparada com a obtida pelo modelo AG), como as estimativas específicas associadas a cada acontecimento k , $\beta_k = (\beta_{k1}, \dots, \beta_{kp})'$. Relativamente ao conjunto de indivíduos em risco, considera-se que estão em risco para o k -ésimo acontecimento, apenas os indivíduos aos quais já foi observado o acontecimento $k - 1$, o que se traduz num conjunto de risco restritivo.

Em ambos os modelos, a construção do intervalo de risco é feita através dos processos de contagem (*counting process*), em que a escala utilizada se refere ao tempo desde o início do estudo, mas onde os tempos até à ocorrência de cada acontecimento têm como instante inicial o instante em que o acontecimento anterior é observado

(acontecimentos instantâneos) ou em que o acontecimento anterior termina (acontecimentos duradouros). Importa salientar que o modelo PWP permite ainda que o intervalo de risco possa ser formulado segundo o tempo por intervalos (*gap time*), onde a escala de tempo diz respeito ao tempo desde o último acontecimento, sendo que neste caso o relógio reinicia a sua contagem voltando ao instante zero após a ocorrência de cada acontecimento.

Nestes dois modelos semiparamétricos, a estimação do vetor de parâmetros de regressão β é feita pelo método da máxima verosimilhança, pelo que se assume que as observações são independentes. Para isso, é preciso adaptar a função de verosimilhança parcial do modelo de Cox ao contexto dos acontecimentos múltiplos. Todavia, é plausível considerar que os tempos associados ao mesmo indivíduo estejam correlacionados entre si, ou seja, que exista correlação intra-individual. Para ter em conta esse facto, foi desenvolvido um estimador robusto da matriz de covariância – estimador *sandwich* – o qual permite efetuar uma correção na estimativa usual da variância e, por conseguinte, averiguar se as observações estão correlacionadas [9]. Mais detalhes sobre esta temática podem ser encontrados em Kelly e Lim [7].

2.2 Modelos paramétricos de Weibull

Conforme citado por Royston e Parmar [14], o sucesso do modelo de Cox fez com que, involuntariamente, seja dedicado pouco esforço ao estudo da função de risco subjacente. Contudo, em determinadas situações, é importante conhecer o comportamento desta função, uma vez que está diretamente relacionada com a forma como a ocorrência dos acontecimentos evolui ao longo do tempo. Além disso, alguns autores como Kwong e Hutton [8] e até o próprio Cox [13], defendem que adotar um modelo paramétrico, quando adequado, pode aumentar a precisão das estimativas dos parâmetros de regressão, o que por sua vez contribui para uma melhor compreensão do fenómeno em estudo. Assim sendo, a adoção de uma abordagem paramétrica pode vir a revelar-se mais apropriada.

Os motivos referidos levaram-nos ao desenvolvimento de dois modelos paramétricos para analisar acontecimentos múltiplos, consoante se considere ou não a estratificação por acontecimento. Tendo por base os modelos de riscos proporcionais e os modelos semiparamétricos AG (2) e PWP (3), a estratégia consistiu em especificar um modelo paramétrico para o tempo. Nesse sentido, optou-se por considerar a distribuição de Weibull com parâmetro de escala $\lambda > 0$ e parâmetro de forma $\gamma > 0$, cuja função de risco pode ser escrita por $h(t) = \lambda \gamma t^{\gamma-1}$, para $t \geq 0$. A razão pela qual se considerou esta distribuição deve-se ao facto de ocupar um lugar de referência na análise de dados de sobrevivência.

Assim sendo, para o i -ésimo indivíduo em estudo, as correspondentes funções de risco dos modelos paramétrico Weibull não estratificado (WNE) e paramétrico Weibull estratificado (WE) são definidas por

$$h(t; \mathbf{z}_{ik}(t)) = \lambda \gamma t^{\gamma-1} \exp(\boldsymbol{\beta}' \mathbf{z}_{ik}(t)), \quad t \geq 0, \quad (4)$$

e

$$h(t; \mathbf{z}_{ik}(t)) = \lambda_k \gamma_k t^{\gamma_k-1} \exp(\boldsymbol{\beta}' \mathbf{z}_{ik}(t)), \quad t \geq 0, \quad (5)$$

em que $\lambda > 0$ e $\gamma > 0$ denotam os parâmetros de escala e forma comuns a todos os acontecimentos, enquanto $\lambda_k > 0$ e $\gamma_k > 0$ representam os parâmetros de escala e forma específicos do acontecimento k , $k = 1, \dots, K$, respetivamente. Assim, o modelo WNE (4) permite analisar situações em que se assume que o risco de ocorrência dos acontecimentos não se altera. Por outro lado, como no modelo WE (5) são considerados parâmetros de escala e forma específicos de cada acontecimento, este modelo abrange situações em que se admite que os acontecimentos têm riscos de ocorrência diferentes.

Nestes dois modelos paramétricos, os métodos de inferência estatística também são baseados na teoria assintótica de máxima verosimilhança. Para estimar os vários parâmetros de cada modelo, admite-se que as observações são censuradas à direita e que os tempos até cada acontecimento e os tempos de censura são independentes. Deste modo, assumindo que a censura é não informativa, para

ambos os modelos a expressão geral da função de verosimilhança é dada por

$$L = \prod_{i=1}^n \prod_{k=1}^K h(t_{ik}; \mathbf{z}_{ik}(t_{ik}))^{\delta_{ik}} S(t_{ik}; \mathbf{z}_{ik}(t_{ik})), \quad (6)$$

onde t_{ik} denota o tempo de observação do indivíduo i referente ao k -ésimo acontecimento, δ_{ik} é a variável que caracteriza o estado do indivíduo em relação ao acontecimento k e $h(t_{ik}; \mathbf{z}_{ik}(t_{ik}))$ e $S(t_{ik}; \mathbf{z}_{ik}(t_{ik}))$ representam a função de risco e a função de sobrevivência associadas ao modelo que estiver a ser ajustado, respetivamente. Embora a construção das funções de verosimilhança dos modelos WNE (4) e WE (5) tenha por base a mesma expressão (6), a função resultante difere consoante a situação, dado que no segundo modelo existe estratificação.

3 Uma aplicação com dados simulados

Com o propósito de exemplificar a aplicação dos quatro modelos formulados anteriormente, recorreu-se ao *software* estatístico R, versão 3.5.0. Embora existam vários conjuntos de dados reais disponíveis neste *software* (por exemplo, em [15]), optou-se por realçar a possibilidade de este poder ser utilizado para simular acontecimentos recorrentes, em particular através do *package* *survsim* [10].

Por conseguinte, foi gerada uma amostra constituída por $n = 1000$ indivíduos, onde se definiu que os mesmos podiam sofrer no máximo $K = 8$ acontecimentos, de forma a ter um número razoável de indivíduos nos estratos respeitantes aos últimos acontecimentos e, assim, evitar que sejam obtidas estimativas muito instáveis. Relativamente ao tempo de *follow-up*, estipulou-se que este seria igual a 1825 dias (o equivalente a 5 anos) pois, em geral, é necessário que o tempo seja suficientemente longo para que o acontecimento possa repetir-se.

Uma vez que o *package survsim* apenas permite que os tempos até cada acontecimento sejam gerados por intermédio das distribuições de Weibull, log-normal ou log-logística, optou-se pela distribuição de Weibull. Para os tempos de censura, acresce ainda a possibilidade de considerar a distribuição uniforme, algo que é usual neste contexto. No entanto, optou-se por escolher a mesma distribuição para ambos os casos. Assim sendo, definiu-se que os tempos seguem distribuições de Weibull com parâmetros de forma distintos, mas maiores do que 1 de modo a que os acontecimentos tivessem riscos de ocorrência diferentes e que as correspondentes funções de risco fossem crescentes.

Por último, o *package survsim* permite que as covariáveis sejam geradas através de três distribuições distintas, tendo-se decidido explorar as três possibilidades. Simulou-se então uma variável discreta com distribuição de Bernoulli com probabilidade de sucesso igual a 0.5, correspondente à covariável x ; uma variável contínua com distribuição uniforme no intervalo $[0, 1]$, correspondente à covariável $x.1$; e uma outra variável contínua com distribuição gaussiana padrão, correspondente à covariável $x.2$. Nesta etapa, estipulou-se que o efeito de cada covariável seria o mesmo em todos os acontecimentos, sendo que os valores fixados foram $\beta_x = 0.6$, $\beta_{x.1} = 0.03$ e $\beta_{x.2} = 0.75$, respetivamente. Importa salientar que o propósito deste trabalho não é efetuar um estudo de simulação, mas sim utilizar a simulação como um meio de obter um conjunto de dados adequado à aplicação dos dois modelos propostos. Assim, a partir deste momento o conjunto de dados simulados será tratado como se fossem dados reais.

Na Tabela 1, sumarizou-se a informação relevante sobre a constituição e evolução do conjunto de indivíduos em risco por acontecimento. Deste modo, obtém-se uma visão global sobre as características dos dados que, posteriormente, pode contribuir para uma melhor compreensão dos resultados obtidos no ajustamento dos modelos. Procedeu-se então à implementação dos modelos, em que para os dois modelos semiparamétricos (AG e PWP) utilizou-se o *package survival* [15] e para os dois modelos paramétricos (WNE e

Tabela 1: Resumo da informação sobre os dados simulados.

	Número do acontecimento							
	1	2	3	4	5	6	7	8
Conjunto de risco	1000	355	143	83	52	27	17	11
Acontecimentos observados	355	143	83	52	27	17	11	8
Percentagem de censura (%)	64.5	59.7	42.0	37.3	48.1	37.0	35.3	27.3

WE) recorreu-se ao *package* **straweb** [4]. Os resultados obtidos no ajustamento de cada modelo encontram-se compilados na Tabela 2. Como é possível constatar, em todos os modelos as covariáveis \mathbf{x} e $\mathbf{x}.2$ têm influência significativa sobre o tempo até a ocorrência do acontecimento e a covariável $\mathbf{x}.1$ não. Ao comparar os valores do parâmetro associado a cada covariável definido no processo de simulação com as respetivas estimativas, observa-se que os modelos PWP e WE são os que melhor conseguem estimar os valores fixados. Na verdade, o modelo PWP é o que apresenta melhores resultados em relação ao efeito das covariáveis \mathbf{x} e $\mathbf{x}.1$, enquanto o modelo WE se destaca em relação ao efeito da covariável $\mathbf{x}.2$.

Considerando a estimativa usual do erro padrão ($EP(\hat{\beta}_j)$), verifica-se que estas são praticamente iguais entre modelos não estratificados (AG e WNE) e modelos estratificados (PWP e WE). Nos modelos semiparamétricos pode-se comparar ainda a estimativa usual do erro padrão com a respetiva estimativa robusta, onde se observa que $EP_r(\hat{\beta}_j)$ é ligeiramente superior a $EP(\hat{\beta}_j)$. Segundo Kelly e Lim [7], este facto está de acordo com a potencial existência de correlação intra-individual.

Em ambas as abordagens, semiparamétrica e paramétrica, o modelo mais adequado para estes dados foi o que considerou a estratificação, o que é coerente com o facto de os dados terem sido simulados de modo a que os acontecimentos tivessem riscos de ocorrência distintos. De facto, através do critério de informação de Akaike (*AIC*), na abordagem semiparamétrica, verifica-se que o modelo

Tabela 2: Estimativas dos parâmetros de regressão para cada modelo.

Covariável/Modelo	$\hat{\beta}_j$	$\exp(\hat{\beta}_j)$	$EP(\hat{\beta}_j)$	$EP_r(\hat{\beta}_j)$	Valor- p
x					
AG	0.806	2.238	0.081	0.107	5.84e-14
WNE	0.829	2.290	0.082	—	< 2e-16
PWP	0.651	1.917	0.083	0.088	1.20e-13
WE	0.683	1.981	0.083	—	2.22e-16
x.1					
AG	0.064	1.066	0.131	0.173	0.714
WNE	0.063	1.065	0.133	—	0.636
PWP	0.048	1.049	0.133	0.137	0.725
WE	0.057	1.059	0.135	—	0.671
x.2					
AG	0.861	2.365	0.046	0.055	< 2e-16
WNE	0.871	2.389	0.045	—	< 2e-16
PWP	0.720	2.055	0.048	0.047	< 2e-16
WE	0.765	2.150	0.048	—	< 2e-16

PWP apresenta um valor AIC inferior ao obtido para o modelo AG ($AIC_{PWP} = 7133.034$ e $AIC_{AG} = 9035.034$) e, na abordagem paramétrica, observa-se que o valor AIC mais baixo encontra-se associado ao modelo WE ($AIC_{WE} = 10825.300$ e $AIC_{WNE} = 11015.160$)⁵.

Na abordagem paramétrica, pode-se ainda avaliar a adequabilidade do modelo, de forma informal, representando graficamente as estimativas da função de sobrevivência obtidas pelo estimador de Kaplan-Meier e pelo modelo nulo⁶ que estiver a ser considerado, tendo-se obtido o gráfico apresentado à esquerda na Figura 1. Através deste gráfico, verifica-se que as estimativas baseadas no modelo de Weibull são bastante próximas das estimativas de Kaplan-Meier,

⁵Note-se que a comparação dos valores AIC obtidos para cada modelo só pode ser feita dentro de cada abordagem, visto que o tipo de verosimilhança considerada não é o mesmo.

⁶O modelo é ajustado sem incluir as covariáveis.

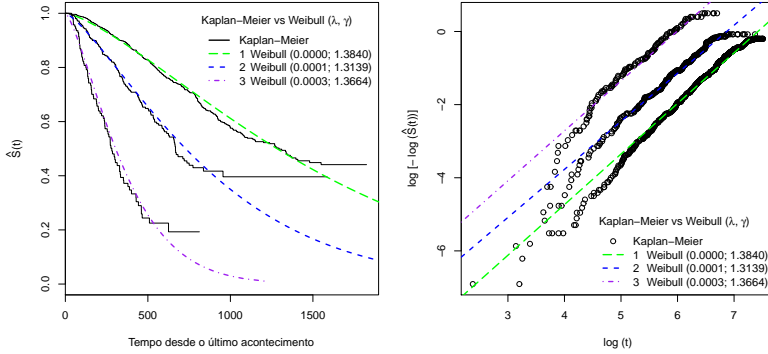


Figura 1: Estimativas da função de sobrevivência de Kaplan-Meier e para a distribuição de Weibull, referentes aos três primeiros acontecimentos.

excepto na parte final. Relembre-se que os dados foram propositalmente simulados através da distribuição de Weibull, pelo que este resultado já seria de esperar. Outra forma de realizar esta análise, consiste em considerar a representação de $\log [-\log (\hat{S}(t))]$ versus $\log(t)$ para cada acontecimento, originando o gráfico apresentado à direita. Analisando este gráfico, confirma-se que de facto assumem uma forma razoavelmente linear, embora tal não aconteça em alguns dos menores tempos observados. Note-se que, para não sobrecarregar o gráfico, decidiu-se apenas representar as estimativas referentes aos três primeiros acontecimentos/estratos.

4 Conclusões e trabalho futuro

De um modo geral, os dois modelos paramétricos baseados na distribuição de Weibull (WNE e WE), revelaram-se uma alternativa bastante adequada às duas extensões do modelo semiparamétrico de Cox que foram propostas para analisar acontecimentos recorrentes,

os modelos AG e PWP.

A implementação dos dois novos modelos foi efetuada com recurso ao *package straweib* [4], originalmente criado para lidar com a violação do pressuposto de riscos proporcionais, permitindo uma função de risco subjacente específica para cada estrato. Assim, o que se fez foi adaptar o código deste *package* ao caso em que são observados vários acontecimentos por indivíduo, utilizando a variável que indica o número do acontecimento como variável de estratificação.

Quando a função de risco subjacente é corretamente especificada, sabe-se que os modelos paramétricos evidenciam maior eficiência comparativamente aos modelos semiparamétricos. Embora a aplicação dos modelos WNE e WE a este conjunto de dados simulados, em particular, não tenha melhorado a precisão das estimativas dos parâmetros de regressão, estes modelos têm a vantagem de permitir estimar de forma suave a função de risco. Além do mais, conhecer a distribuição associada ao tempo até cada acontecimento contribui para uma melhor compreensão sobre o modo como a multiplicidade de acontecimentos evolui ao longo do tempo.

A abordagem aplicada neste trabalho consistiu em considerar a distribuição de Weibull para especificar a forma da função de risco subjacente e, conseqüentemente, obter um modelo totalmente paramétrico. Porém, consoante o caso de estudo, tem interesse ponderar outras distribuições que podem vir a revelar-se mais apropriadas ou até mesmo ter maior flexibilidade para captar a forma como o risco evolui ao longo do tempo. Nesse sentido, recomenda-se explorar o *package flexsurv* [6] para implementar esses novos modelos.

Apesar de esta investigação ter sido direcionada para acontecimentos recorrentes, isso não significa que estes dois modelos paramétricos não possam ser utilizados para analisar acontecimentos de tipos diferentes. Uma outra temática a explorar futuramente seria desenvolver um estimador robusto para a matriz de covariância, à semelhança daquele que foi desenvolvido para as extensões do modelo de Cox.

Agradecimentos

O primeiro autor agradece à Sociedade Portuguesa de Estatística (SPE), pela bolsa que lhe foi concedida para participar no “XXIII Congresso da SPE”. Esta investigação foi parcialmente financiada por Fundos Nacionais, através da Fundação para a Ciência e a Tecnologia (FCT), no âmbito dos projetos UID/MAT/00006/2013 (Centro de Estatística e Aplicações) e UID/MAT/04674/2013 (Centro de Investigação em Matemática e Aplicações).

Referências

- [1] Andersen, P. K. e Gill, R. D. (1982). Cox’s regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4), 1100–1120.
- [2] Cox, D. R. (1972) Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34(2), 187–220.
- [3] Ferreira, I. M. S. (2016). *Modelos para acontecimentos múltiplos*. (Dissertação de Mestrado). Universidade da Madeira, Funchal, Portugal.
- [4] Gu, X. e Balasubramanian, R. (2013). straweib: Stratified Weibull Regression Model. *Package do R versão 1.0*. URL: <https://github.com/cran/straweib>.
- [5] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York.
- [6] Jackson, C., Metcalfe, P. e Amdahl, J. (2017). flexsurv: Flexible Parametric Survival and Multi-State Models. *Package do R versão 1.1*. URL: <https://CRAN.R-project.org/package=flexsurv>.
- [7] Kelly, P. J. e Lim, L. L-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine*, 19(1), 13–33.
- [8] Kwong, G. P. S. e Hutton, J. L. (2003). Choice of parametric models in survival analysis: applications to monotherapy for epilepsy and cerebral palsy. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(2), 153–168.

- [9] Lin, D. Y. e Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84(408), 1074–1078.
- [10] Moriña, D. e Navarro, A. (2015). *survsim: Simulation of Simple and Complex Survival Data*. Package do R versão 1.1.5. URL: <https://CRAN.R-project.org/package=survsim>.
- [11] Prentice, R. L., Williams, B. J. e Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68(2), 373–379.
- [12] R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- [13] Reid, N. (1994). A conversation with Sir David Cox. *Statistical Science*, 9(3), 439–455.
- [14] Royston, P. e Parmar, M. K. B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15), 2175–2197.
- [15] Therneau, T. M. (2015). *survival: A package for survival analysis in S*. Package do R versão 2.41-3. URL: <https://CRAN.R-project.org/package=survival>.
- [16] Therneau, T. M. e Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Modelagem de capturas em peso inflacionadas de zeros no Baixo Rio Amazonas

Júlio César Pereira

Departamento de Ciências Ambientais, Universidade Federal de São Carlos, Rodovia João Leme dos Santos, KM 110, Sorocaba-SP, Brasil, *julio-pereira@ufscar.br*

Giovani L. Silva

CEAUL & Dep. de Matemática, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal, *giovani.silva@tecnico.ulisboa.pt*

Victória Isaac

Laboratório de Biologia Pesqueira e Manejo de Recursos Aquáticos, Universidade Federal do Pará, Brasil, *biologiapesqueira@hotmail.com*

Palavras-chave: Inflação de zeros; Pesca; Poisson composta; Modelagem hierárquica.

Resumo: Este trabalho visa aplicar modelos estatísticos padrão para analisar dados de captura por unidade de esforço da pesca de pequena escala e, em particular, do Baixo Rio Amazonas. Neste sentido, desenvolvemos um modelo hierárquico bayesiano em três etapas. Na primeira etapa, descrevemos o número de viagens por local de pesca (N) de acordo com a distribuição de Poisson, enquanto na segunda etapa, dado $N > 0$, definimos uma variável de Bernoulli X com probabilidade q de haver captura de uma dada espécie. Na terceira etapa, modelamos o peso Y de pesca capturada, com $Y = 0$, quando $N = 0$ ou $X = 0$ e $N > 0$, e $Y > 0$, no caso contrário. Quando $X = 1$ e $N > 0$, descrevemos Y de acordo com a distribuição gama. O modelo proposto pretende ser uma ferramenta útil para analisar a variação na captura por unidade de esforço em função de

covariáveis, tendo em conta que os dados podem estar inflacionados de zeros, provenientes de ambas as fontes: abstinência da atividade pesqueira ou ausência de captura na presença de atividade pesqueira.

1 Introdução

Dados de captura e esforço gerados por pescarias comerciais são comumente disponibilizados para avaliar estoques pesqueiros, juntamente com covariáveis relacionadas ao tipo de embarcações e ao ambiente de pesca. Neste caso, a captura por unidade de esforço (CPUE) é usualmente usada para designar a quantidade efetiva de pesca, podendo ser *e.g.* a biomassa de peixes capturados por número de pescadores e dia de pesca. Geralmente a quantidade CPUE é utilizada como um índice de abundância, por isso, ao se fazer a avaliação de um estoque pesqueiro há interesse em detetar tendências nesse índice ao longo do tempo. Por um lado, como dados de CPUE nominais podem conter efeitos de covariáveis, modelos estatísticos são utilizados na padronização das CPUE, de forma que essas quantidades ajustadas para um mesmo nível de covariáveis sejam consideradas como índice de abundância. Por vezes o esforço de pesca também pode ser modelado ([13]). Por outro lado, dificuldades no ajuste de modelos para a CPUE ou para a captura surgem quando há elevada proporção de zeros nessas variáveis.

Algumas abordagens têm sido propostas na literatura para a modelagem de dados de CPUE ou de captura na presença de zeros. Mais comumente, os modelos lineares generalizados com distribuição Poisson ou binomial negativa, quando se observa a captura em número de indivíduos. Entretanto, quando a proporção de zeros nas contagens é maior do que a esperada sob uma distribuição Poisson ou binomial negativa, adotam-se os chamados modelos inflacionados de zeros [4] ou alternativamente os modelos com barreira (*hurdle*) (*vide e.g.* [10]).

No contexto de modelo contínuo, a modelagem da CPUE ou captura em peso, com elevada proporção de zeros, é por vezes feita via mo-

delos do tipo “delta”. Nesse caso, a probabilidade de ocorrência de zeros pode ser descrita pelo modelo logístico, enquanto que para o total de capturas se considera geralmente o modelo log-normal, gama ou gaussiana inversa. Alternativamente, pode-se usar aqui modelos com distribuição composta (*vide e.g.* [1, 11, 5, 6]), tal como a distribuição Poisson-gama, que pode ser vista como caso particular da chamada família de distribuições de Tweedie [3].

Nos modelos Poisson compostos, as capturas em peso ou CPUE, denotas por Y_i , são considerada resultantes de uma soma de capturas em peso (ou CPUE) individuais W_1, W_2, \dots, W_{N_i} , onde os W_k , $k = 1, \dots, N$, são variáveis aleatórias independentes e identicamente distribuídas de acordo com distribuição gama e N_i segue uma distribuição Poisson. A variável N representa *e.g.* o número de indivíduos por lance de pesca, sendo W o peso de cada indivíduo associado [1]. Alternativamente N pode ser o numero de aglomerados de peixes encontrados e W o peso obtido em cada aglomerado [5]. Como N pode ser zero, com probabilidade determinada pela distribuição de Poisson, então a distribuição de Y tem uma massa de probabilidade maior que zero em $Y = 0$. Dessa forma, essa abordagem tem a vantagem de modelar de forma unificada o peso individual e a presença/ausência de capturas.

Nessa abordagem observa-se que as quantidades Y associadas a $N > 0$ são sempre positivas. Entretanto, se N representar *e.g.* o número de viagens de pesca ou o número de lances e Y for a quantidade produzida em cada viagem, então é possível ter $N > 0$ com o respetivo $Y = 0$. Nessa situação, o modelo Poisson composto já não é aplicável, pois não acomoda os valores de captura zeros associados a $N > 0$. Essa situação motivou o desenvolvimento do presente trabalho.

Ao analisar dados de captura e esforço de pescarias no Baixo Rio Amazonas deparamos com a dificuldade em aplicar os modelos inicialmente mencionados, uma vez que a produção em peso de capturas mensais por localização e por espécie apresentavam zeros que podiam ocorrer por dois motivos: i) porque as localizações não eram visitadas por qualquer embarcação naquele mês ou ii) porque havia

mais que uma visita, mas para parte das visitas não havia sucesso na captura da espécie de interesse.

O objetivo do presente trabalho é desenvolver um modelo capaz de acomodar essas duas fontes de zeros na captura em peso, ou seja, captura zero resultante de $N = 0$, a exemplo do modelo Poisson Composto, bem como captura zero devido ao fracasso na captura quando $N > 0$.

2 Dados

Os dados analisados no presente trabalho são provenientes de pescarias comerciais ocorridas no Baixo Rio Amazonas e desembarcadas no ano de 2004 no porto de Santarém-Pará, o maior porto da região. As variáveis registradas são i) o número de pescadores na embarcação, ii) o número de dias efetivos de pesca, iii) a época de pesca que foi dividida em dois períodos: *mo* - período de março à outubro, que vai do fim da cheia e início da seca; *nf* - período de novembro à fevereiro, o qual abrange as estações seca e enchente, iv) o ambiente de pesca, o qual foi dividido em duas categorias: rio - quando a pesca ocorreu no canal principal ou canais menores do rio - e lago - quando a pesca ocorreu em lagos de várzea ou áreas alagadas lateralmente. O peso (em Kg) das capturas de cada espécie desembarcada também foram registrados.

A Figura 1 ilustra as localizações das atividades de pesca. Para a modelagem das capturas neste estudo escolheu-se a espécie Mapará (*Hypophthalmus marginatus* and *H. edentatus*) por ser a espécie que apresentou em geral a maior captura em peso. Mais detalhes dos dados podem ser obtidos em [9]. A variável esforço de pesca foi definida em termos do produto do número de pescadores por dias efetivos de pesca. Havendo agregação de capturas por localização e mês, o esforço também teve essa agregação.

O conjunto de dados é formado por 393 unidades amostrais, das quais 126 são caracterizadas por número de visitas igual a zero. E dentre as 267 unidades para as quais houve visitas, em 202 não houve

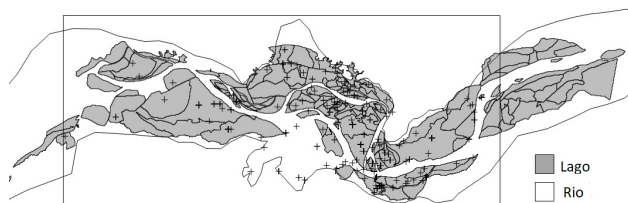


Figura 1: Localização de atividades pesqueiras, identificada por “+”.

sucesso na pesca do Mapará. A Figura 2 apresenta o histograma do peso capturado (em toneladas) incluindo-se capturas iguais a zero, ilustradas por um segmento fracionado nas cores cinza ($N = 0$) e preto ($Y = 0|N > 0$), com média de 477 Kg (desvio padrão de 2158) para todas as capturas e média 800 Kg (desvio padrão de 4635) considerando apenas capturas positivas.

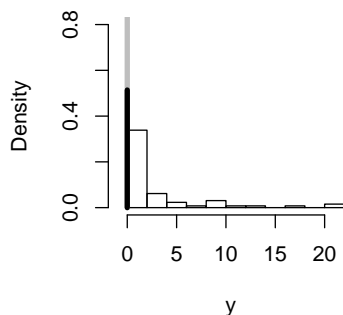


Figura 2: Histograma do peso capturado do Mapará por local-mês.

3 Modelos

Para a análise do conjunto de dados em estudo, desenvolvemos um modelo hierárquico bayesiano em três etapas. Na primeira etapa, descrevemos o número de viagens de pesca (N) por mês e localização

de acordo com uma distribuição de Poisson

$$N_i \sim \text{Poisson}(\lambda_i), \quad i = 1, 2, \dots, n, \quad (1)$$

em que n representa o número de unidades amostrais. A média de N_i , representada por λ_i foi considerada dependente das covariáveis disponíveis, *i.e.*,

$$\log \lambda_i = \beta_0 + \beta_1 \text{rio}_i + \beta_2 \text{mo}_i, \quad (2)$$

onde *rio* e *mo* são variáveis “dummies”, com *rio* = 1 (pesca em ambiente de rio) e *rio* = 0 (pesca em ambiente de lago), *mo* = 1 (pesca de março a outubro) e *mo* = 0 (pesca de novembro a fevereiro). A fim de modelar possível sobredispersão nas contagens, também consideramos uma extensão do modelo (2), isto é, $\log \lambda_i = \beta_0 + \beta_1 \text{rio}_i + \beta_2 \text{mo}_i + v_i$, onde v_i é um efeito aleatório, com distribuição gaussiana de média zero e variância σ^2 .

Na segunda etapa, dado que o esforço foi maior que zero para a observação i , definimos uma variável Bernoulli X com probabilidade q de sucesso, onde X é igual a 1, se a captura ocorre para a espécie de interesse, e zero, no caso contrário. Isto é,

$$X_i \mid N = n_i \sim \text{Bern}(q_i), \quad \forall n_i > 0, \quad (3)$$

em que $q_i = P(X_i = 1 \mid N = n_i)$, potencialmente dependente das covariáveis disponíveis, tem uma estrutura dada por

$$\text{logit}(q_i) \equiv \log [q_i / (1 - q_i)] = \mathbf{Z}_i \boldsymbol{\gamma}, \quad (4)$$

onde \mathbf{Z}_i é um vetor linha de covariáveis observadas, possivelmente com dimensão diferente dos outros vetores de regressão a referir aqui *e.g.* associado à média λ_i . Em princípio, ajustou-se o modelo incluindo-se as covariáveis *rio* e *mo* a fim de selecionarmos posteriormente aquelas que eram significativas. Uma das versões dos modelos ajustados incluiu também o esforço de pesca como um *offset* na equação do preditor linear, visto que o esforço de pesca está diretamente ligado à captura.

Na terceira etapa, modelamos o peso Y , que assume zero quando $N = 0$ ou quando $N > 0$ e $X = 0$. Quando $N > 0$ e $X = 1$ descrevemos Y de acordo com uma distribuição gama. Assim, o i -ésimo peso capturado é dado por

$$Y_i = \begin{cases} 0 & \text{se } N_i = 0 \\ 0 & \text{se } N_i > 0 \text{ e } X = 0 \\ \sum_{k=1}^{N_i} W_k & \text{se } N_i > 0 \text{ e } X_i = 1 \end{cases} \quad (5)$$

onde os pesos capturados W_k para cada evento de pesca bem sucedido são considerados independentes e identicamente distribuídas, seguindo distribuição gama, denotada por $W_k \sim Gama(a_0, b)$, portanto o peso total dado $N_i = n_i, n_i > 0$ e $X_i = 1$ também segue distribuição gama, $Y_i | N_i = n_i, X_i = 1 \sim Gama(a_0 \times N_i, b)$ [2].

A partir de uma análise exploratória dos dados sugerindo uma relação linear entre o logaritmo dos pesos capturados e o número de viagens de pesca, assumimos que $Y_i | N_i = n_i, X_i = 1 \sim Gama(a, b)$ com o logaritmo da média proporcional a N_i . Dessa forma, em princípio incluímos também as covariáveis disponíveis (*rio* e *mo*) na média da distribuição gama. Isto é, a média da distribuição gama inicialmente ficou definida como

$$E[Y_i | N_i = n_i, X_i = 1] = \frac{a}{b} = e^{(m_0 + m_1 \times rio + m_2 \times mo_i + N_i)}, \quad (6)$$

obtendo-se $V[Y_i | N_i = n_i, X_i = 1] = \frac{e^{2(m_0 + m_1 \times rio + m_2 \times mo_i + N_i)}}{a}$. Novamente, em uma das versões dos modelos ajustados incluímos o esforço de pesca como um *offset* na média do peso capturado.

3.1 Inferência

Sob uma abordagem bayesiana, a especificação completa do modelo acima requer distribuições *a priori* para os parâmetros. Ou seja, para os parâmetros β em (2), γ em (4) e $\mathbf{m} = (m_0, m_1, m_2)$ foram consideradas distribuições *a priori* normais vagas, tendo em conta a falta de informação *a priori* no estudo vigente: $\beta \sim N(\mathbf{0}, 100\mathbf{I})$,

$\mathbf{m} \sim N(\mathbf{0}, 100\mathbf{I})$ e $\gamma \sim N(\mathbf{0}, 100\mathbf{I})$, onde \mathbf{I} é uma matriz identidade $p \times p$ e p é a dimensão do respetivo vetor de parâmetros.

Nota-se que é necessário atribuir distribuição *a priori* apenas a um dos parâmetros a ou b , associados ao vetor \mathbf{m} da equação 6. Por conveniência computacional atribuímos distribuição para o parâmetro a , isto é, distribuição *a priori* vaga, gama($a_1 = 0.01, b_1 = 0.01$), garantindo uma melhor implementação da distribuição em causa.

As inferências sobre os parâmetros de interesse baseiam-se nas suas distribuições *a posteriori*. Supondo independência *a priori*, a distribuição conjunta *a posteriori* é expressa por

$$\pi(\boldsymbol{\theta}; \mathbf{y}, \mathbf{N}, \mathbf{x}) \propto L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{N}, \mathbf{x}) \times \pi(\boldsymbol{\beta}) \times \pi(\boldsymbol{\gamma}) \times \pi(\mathbf{m}) \times \pi(b) \quad (7)$$

em que $L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{N}, \mathbf{x})$ representa a função de verosimilhança do modelo em causa e $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{m}, b)$. Infelizmente, a distribuição *a posteriori* (7) não tem uma apresentação analítica razoável e portanto usamos métodos de Monte Carlo via cadeias de Markov (MCMC) para gerar amostras das distribuições marginais *a posteriori* dos parâmetros, utilizando o programa OpenBugs [7], com as 200 mil primeiras amostras descartadas (*burn-in*) e espaçamento igual a 100 (*emphthin*) e amostra final de 1500.

3.2 Seleção de modelos

Para a seleção de um modelo hierárquico bayesiano em três etapas, descrito na Secção 3, usaram-se os critérios de comparação de modelos: DIC (*Deviance Information Criterion*) [12] e WAIC (*Watanabe-Akaike Information Criteria*)[14]. Modelos com menores valores de DIC e WAIC são os preferidos na seleção. Além disso, descartaram-se inicialmente as covariáveis em cada etapa do modelo para as quais os respetivos intervalos de credibilidade a 95% continham o valor zero.

Foram também utilizados como medidas de diagnóstico e adequabilidade dos modelos a soma dos logaritmos das densidades preditivas de cada observação y_i : i) condicionalmente aos dados \mathbf{y}_{-i} i.e. deixando a i^{th} observação fora do ajuste (LCPO), ii) condicionalmente

aos dados completos \mathbf{y} (Lppd). Note-se que, quanto maiores forem os valores de Lppd e LCPO, tanto mais adequado será um modelo.

4 Resultados

A Tabela 1 apresenta a definição dos 7 modelos ajustados, com a presença (\checkmark) ou não de parâmetros de regressão, parâmetro de dispersão σ^2 , *offset* esforço de pesca (*eff*) e parâmetro de forma a associados às quantidades λ , q e μ em cada etapa do modelo. Nota-se que quando uma célula está vazia, significa que o respetivo parâmetro não faz parte do respetivo modelo. Além disso, os efeitos de covariáveis “significativos”, com base em intervalos de credibilidade a 95%, são denotados por \star , optandos-se pela sua omissão por limitação de espaço. Por exemplo, como o zero está contido no intervalo de credibilidade (omitido aqui por limitação de espaço) para o parâmetro m_1 no modelo M1, deve-se excluir a covariável *rio* associada a esse parâmetro para μ , dando origem aqui ao modelo M2. Esse resultado sugere que as covariáveis *rio* e *mo* afetam o número de viagens e a probabilidade de captura e que apenas a variável *mo* afeta as intensidades das capturas.

A seguir à seleção preliminar de modelos, os vários modelos hierárquicos para a captura do Mapará, usaram-se critérios de comparação de modelo, descritos anteriormente e apresentados agora na Tabela 2 para os modelos M2, M4, M5 e M7. Esses critérios são apresentados apenas para as versões dos modelos em que os intervalos de credibilidade 95% de todos os parâmetros não contém o valor zero, igualmente confirmados pelos respectivos valores das medidas de comparação dos modelos.

Comparando-se o modelo M4 ao modelo M2 nota-se uma redução no valor de DIC e pequena diferença nos valores de WAIC, bem como nos valores de Lppd e LCPO. Esperava-se um ganho nas predições do M4 com relação ao modelo M2, por se acrescentar o *offset* para explicar o sucesso/fracasso nas capturas, dado a presença de pesca. De fato, a proporção de vezes que se conseguiu prever corretamente

Tabela 1: Seleção preliminar de modelos com base em intervalos de credibilidade a 95% dos parâmetros dos modelos.

		M1	M2	M3	M4	M5	M6	M7
λ :	β_0	✓, *	✓, *	✓, *	✓, *	✓, *	✓	
	β_1	✓, *	✓, *	✓, *	✓, *	✓, *	✓, *	✓, *
	β_2	✓, *	✓, *	✓, *	✓, *	✓, *	✓, *	✓, *
	σ^2						✓	✓
q :	γ_0	✓, *	✓, *	✓, *	✓, *	✓, *	✓, *	✓, *
	γ_1	✓	✓, *	✓				
	γ_2	✓, *	✓, *	✓, *	✓, *	✓, *	✓, *	✓, *
	<i>eff</i>			✓	✓	✓	✓	✓
μ :	m_0	✓, *	✓, *	✓, *	✓, *	✓	✓, *	✓, *
	m_1	✓						
	m_2	✓, *	✓, *	✓, *	✓, *	✓	✓, *	✓, *
	a	✓	✓	✓	✓	✓	✓	✓
	<i>eff</i>					✓		

o fracasso e o sucesso nas capturas aumentaram de 0.80 para 0.93 e de 0.28 para 0.66 respetivamente quando comparado ao modelo M2 (Tabela 3). Quanto ao modelo M5, a inclusão do *offset* também na média do peso capturado, resultou num aumento drástico dos valores de DIC e WAIC e redução nos valores de Lppd e LCPO (Tabela 2) sem no entanto se ter melhoras nas predições do fracasso e sucesso das capturas (Tabela 3). Esses resultados sugerem que não há melhora no ajuste e na capacidade preditiva do modelo ao se incorporar o *offset* na equação da média do peso capturado. Cabe notar que o número de viagens é uma componente da média do peso capturado, sendo assim, a inclusão do esforço não contribui para a melhora do modelo possivelmente pela relação de dependência entre essas duas variáveis.

O modelo M7 inclui uma componente aleatória no preditor linear do número de viagens N . Comparado ao modelo M4, esse modelo apresentou redução no valor de DIC e um pequeno aumento no valor

Tabela 2: Seleção final de modelos.

Model	<i>DIC</i>	<i>WAIC</i>	Lppd	LCPO
M2	-1118.51	1048.31	-518.79	-524.18
M4	-1384.63	1050.16	-518.93	-525.27
M5	1626.14	1637.33	-808.30	-818.77
M7	-1819.76	1078.09	-489.27	-541.00

Tabela 3: Proporção de predições corretas de zero, n° de visitas (positivas), capturas nulas e capturas positivas, dado que as visitas ocorreram, com base nos modelos propostos.

Model	$N = 0$	$N > 0$	$Y = 0 \mid N > 0$	$Y > 0 \mid N > 0$
M2	0.26	0.89	0.80	0.28
M4	0.29	0.91	0.93	0.66
M5	0.29	0.89	0.93	0.66
M7	1	1	0.91	0.68

de WAIC, bem como aumento no Lppd e redução no LCPO. Uma vantagem desse modelo com relação a todos os demais foi com relação à sua capacidade de prever corretamente valores de N iguais a zero. Esse modelo prediz corretamente valores $N = 0$ em 100% dos casos, enquanto que essa percentagem foi de no máximo 33% para os modelos concorrentes (Tabela 3). Esse é um indicativo de que extra variação está presente nas contagens e que essa variação não deve ser ignorada.

Usando a mediana *a posteriori* como um estimador pontual para os parâmetros do modelo selecionado (M7), apresentamos na Tabela 4 estimativas do número esperado de visitas, da probabilidade de sucesso nas capturas dado a presença de visitas (q), bem como das probabilidades de capturas iguais a zero para as diversas situações. As medianas dos parâmetros do modelo selecionado (Modelo M7) foram $\hat{\beta}_1 = -1.26$, $\hat{\beta}_2 = 0.43$, $\hat{\gamma}_0 = -15.81$, $\hat{\gamma}_2 = 7.6$, $\hat{m}_0 = -3.26$, $\hat{m}_2 = 3.58$, $\hat{a} = 0.61$, $\hat{\sigma}_2 = 1.67$. O esforço foi fixado em 110 unidades para o cálculo apresentado.

Tabela 4: Valores estimados do numero médio de visitas, da probabilidade de sucesso nas capturas dado a presença de pesca e da probabilidade de captura igual a zero para diferentes cenários de Ambiente (Amb.) e Período (Per.), com esforço de 110 unidades.

Amb. e Per.	$\hat{\lambda}$	q	$Pr(N = 0)$	$Pr(Y = 0 \mid N > 0)$	$Pr(Y = 0)$
Lago and nf	1.00	0.01	0.37	0.63	0.99
Lago and mo	1.54	0.94	0.21	0.05	0.26
Rio and nf	0.28	0.01	0.75	0.24	0.99
Rio and mo	0.44	0.94	0.65	0.02	0.67

No ambiente *lago* e durante o período março a outubro é esperado o maior número de visitas (maior valor de $\hat{\lambda}$) enquanto que no ambiente *rio* e nos meses de novembro a fevereiro o menor número de visitas é esperado. De acordo com o modelo selecionado, a maior chance de captura (q) dada a presença de visitas de pesca ocorre no período de março a outubro. A probabilidade de sucesso na captura não depende do ambiente, entretanto depende do esforço de pesca, que está relacionado ao número de visitas e que por sua vez depende do ambiente.

5 Discussão

De forma semelhante à proposta de [5] e [11], utilizamos uma modelagem hierárquica, em que tínhamos a variável peso capturado associado a contagens, dada pelo número de visitas de embarcações de pesca. Nas referências citadas as observações iguais a zero na variável contínua eram resultantes de contagens iguais a zero. No presente artigo expandimos esse modelo, ao substituir a distribuição gama por uma distribuição delta-gamma flexibilizamos o modelo e permitimos que as capturas iguais a zero pudessem ser devido às contagens iguais a zero bem como devido ao fato de se ter número de visitas maior que zero associado a fracasso na captura. Ao permi-

tir covariáveis nas três etapas do modelo, o modelo proposto ainda permite identificar os níveis das covariáveis que proporcionam maior chance de captura zero devido ao número de visitas igual a zero bem como devido ao fracasso nas capturas.

A introdução de uma componente aleatória no preditor linear de N também se mostrou importante. Os modelos sem a componente aleatória, a exemplo do modelo M4, não permitia a extração de zeros da distribuição Poisson que fosse compatível à quantidade de zeros dos dados. Mesmo quando a estimativa da média da Poisson era pequena, não se obtinha muitos zeros, pelo fato da variância também ser pequena. Quando se incluiu variância extra, pela inclusão dessa variável latente, flexibilizou-se a distribuição Poisson, isto é, a variância não mais ficou vinculada ao valor da média, e portanto, quando a média era próxima de zero mais amostras iguais a zero eram extraídas.

De forma geral, o modelo aqui proposto, mostrou-se flexível e útil para analisar a variação na captura como uma função de covariáveis quando os dados foram inflacionados de zeros provenientes de ambas as fontes: abstinência da atividade pesqueira e ausência de captura na presença de atividade pesqueira. Uma limitação deste trabalho é a não inclusão de uma componente espaço-temporal que se espera ser investigada em trabalhos futuros, bem como uma análise de sensibilidade cabal sobre os parâmetros de forma e dispersão.

Referências

- [1] Candy, S.G. (2004). Modeling catch and effort data using generalized linear models, the Tweedie distribution, random vessel effects and random stratum-by-year effects. *CCAMLR Science*, 11, 59–80
- [2] Hogg, R.V., McKean, J.W., Craig, A.T. (2004). *Introduction to Mathematical Statistics* (6th ed.). USR, New Jersey: Prentice Hall.
- [3] Jorgensen, B., Souza, M.C.P. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scand. Actuar. J.*, 1, 69–93.
- [4] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1–14.

- [5] Lecomte, J.B., Benoit, H.P., Ancelet, S., Etienne, M.P., Bel, L., Parent, E. (2013). Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling volume. *Methods in Ecology and Evolution*, 4, 1159–1166.
- [6] Lecomte, J.B., Benoit, H.P., Etienne, M.P., Bel, L., Parent, E. (2013). Modeling the habitat associations and spatial distribution of benthic macroinvertebrates: A hierarchical Bayesian model for zero-inflated biomass data. *Ecological Modelling*, 265, 74–84.
- [7] Lunn, D., Spiegelhalter, D., Thomas, A., Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Stat. Med.*, 28, 3049–3067.
- [8] Minami, M., Lennert-Cody, C.E., Gao, W., Roman-Verdesoto, M, (2007). Modeling shark bycatch: The zero-inflated negative binomial regression model with smoothing. *Fisheries Research*, 84, 210–221.
- [9] Pinaya, W.H.D., Lobon-Cervia, F.J., Pita, P., Buss de Souza, R., Freire, J., Isaac, V.J. (2016). Multispecies fisheries in the Lower Amazon River and its relationship with the regional and global climate variability. *PLoS ONE*, 11, e0157050.
- [10] Ridout, M., Demetrio, C.G.B., Hinde, J. (1998). *Models for Count Data with Many Zeros*. International Biometric Conference.
- [11] Shono, H. (2008). Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research*, 93, 154–162.
- [12] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. (2002). Bayesian measures of model complexity and fit. *J. Royal Stat. Society B*, 64, 583–639.
- [13] Vermard, Y., Rivot, E., Mahévas, S., Marchal, P., Gascuel, D. (2010) Identifying fishing trip behaviour and estimating fishing effort from VMS data using Bayesian Hidden Markov Models, *Ecological Modelling*. 221, 15, 1757–1769.
- [14] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571–3594.

Optimal re-sampled efficient frontier and examples

Marcos Huber Mendes

Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro,RJ,
Brasil, *hubermendes@decisionsupport.com.br*

Reinaldo Castro Sousa

Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro,RJ,
Brasil, *reinaldo@puc-rio.br*

Marco Aurélio Sanfins

Universidade Federal Fluminense, Niterói,RJ,
Brasil, *marcosanfins@automata.uff.br*

Keywords: Optimal; Modal; Re-Sampled; Frequency; Efficient Frontier.

Abstract: This paper presents contributions to the generalization of Markowitz's Portfolio Optimization Model [11]. First, we define a new risk measure considering all cross interrelationships between returns in addition to deviations above and below a target. Second, we introduce frequency analysis to Re-Sampled Portfolio Assets in order to consider the distribution frequency of the optimal results. This allows to evaluate the Modal Optimum and the Modal Re-Sampled Efficient Frontier, giving a probability measure of the optimal occurrence. The new risk measure uses a quadratic optimization (QP), with a global optimum and lower computational than non-smooth optimization (NSP) used on partial moments and target deviations risks measure.

1 Efficient Frontier Model

1.1 E-V Efficient Portfolio (Markowitz, 1991)

The mathematical formulation of the E-V (Mean,Variance) Efficient Portfolio [10] [11] model (EV) considers three basic premises: i) uncovered sales are not allowed; ii) the sum of the fractions to invest equals the capital available for investment; and iii) the Assets are correlated but not perfectly correlated, implying that diversification can reduce but not eliminate risk. Return is measured as the mean and risk as the variance of the returns of the Portfolio of Assets. Based on these assumptions, the EV optimization model seeks to find the Efficient Portfolio that offers lowest risk for each level of predefined minimum required returns (MRR). The set of Efficient Portfolios for different levels of MMR constitutes the Efficient Frontier (EF), which limits the feasible Portfolios. The EV optimization model is a quadratic model (QP) which allows a global optimum.

1.2 The Mean-separated Target Deviations

The Mean-separated Target Deviations (MSTD) model [8] , unlike the EV model, considers the risk as a joint measure of deviation below (BTD) and above (ATD) a certain target, the Minimum Acceptable Return (MAR), by using the concept of non-central semi-moment or deviation around a target. The MSTD risk measure is a generalization of the risk metrics evolution of the EV model risk measures. These evolution use the concept of non-central semimoment and the concept of lower partial moments (LPM) for risk, and the return as a function of UPM [6]. In particular, the MSTD model can be reduced to the semi-variance above and below a target, the risk measure proposed as in [4], to the risk measure as in [11], and to the risk measure as in [1]. However, the semi-moment of order 2 for a target does not assume a quadratic form, which prevents construction of the Portfolio semi-moment from the n semi-moments and co-variate semi-moments of order 2. In general, for different or-

ders degrees, one does not have a literal form that allows simplifying the computational complexity of a discontinuous and nonlinear algorithm for solving semi-moment models. This means that we have to use empirical data to solve the optimization model algorithm, without a literal form expression, which for that model makes the optimization algorithm a NSP model with a complex solution and high computational consumption and which only provides a single viable solution or a local optimum.

2 The New Statistic Efficient (SE) Risk Measure

Define M as the matrix of the integral of squared deviations and cross deviations of Portfolio returns to a target t:

$$\int_{-\infty}^{\infty} (x_i - t)(x_j - t) f_{ij}(x_i, x_j) dx_i dx_j \quad (1)$$

where t is the target, or MAR, and f_{ij} represents the joint density probability function of the Assets i and j. Assuming $i = j$ in (1), we obtain:

$$M(i, i) = \int_{-\infty}^{\infty} (x_i - t)^2 f_i(x_i) dx_i$$

Proposition 1: The matrix M can be decomposed into the sum of $2n^2 - 2n + 2$ matrices, with $1 \leq i \leq j \leq n$, denoted by:

$$M^+(i, i), M^-(i, i) \text{ and } M_{ij}^{++}(i, j), M_{ij}^{+-}(i, j), M_{ij}^{-+}(i, j), M_{ij}^{--}(i, j),$$

based on the returns of Assets and on the target t, such that:

$$M^+(i, i) = \int_t^{\infty} (x_i - t)^2 f_i(x_i) dx_i \text{ and } M^+(i, j) = 0, i \neq j,$$

$$M^-(i, i) = \int_{-\infty}^t (x_i - t)^2 f_i(x_i) dx_i \text{ and } M^-(i, j) = 0, i \neq j,$$

$$M_{ij}^{++}(i, j) = M_{ji}^{++}(j, i) = \int_t^{\infty} \int_t^{\infty} (x_i - t)(x_j - t) f_{ij}(x_i, x_j) dx_i dx_j \text{ and 0 for the others entries,}$$

$M_{ij}^{+-}(i,j)=M_{ji}^{+-}(j,i)=\int_{x_j=-\infty}^t \int_{x_i=t}^{\infty} (x_i-t)(x_j-t)f_{ij}(x_i,x_j)dx_i dx_j$ and 0 for the others entries,

$M_{ij}^{-+}(i,j)=M_{ji}^{-+}(j,i)=\int_{x_j=t}^{\infty} \int_{x_i=-\infty}^t (x_i-t)(x_j-t)f_{ij}(x_i,x_j)dx_i dx_j$ and 0 for the others entries,

$M_{ij}^{--}(i,j)=M_{ji}^{--}(j,i)=\int_{x_j=-\infty}^t \int_{x_i=-\infty}^t (x_i-t)(x_j-t)f_{ij}(x_i,x_j)dx_i dx_j$ and 0 for the others entries.

The properties of matrices and the split of integrals easily perform the deduction of this result.

We construct the proposed risk measure, referred to as Statistic Efficient (SE) risk measure, based on the empirical formulas analogous to the matrix M. Since M, in addition to taking into account the deviations above and below the target t (MAR), considers all inter-relationships between these deviations, SE also allows the solution of the optimization model through a literal form based on the returns of the Assets and the target t .

Proposition 2: The squared deviation from the return r_{ps} on a Portfolio of Assets A_1, A_2, \dots, A_n related to a MRR t can be written as a function of squared deviations and cross squared deviations of the returns x_i of each Asset A_i and fraction to invest w_i .

Proof: Indeed,

$$\begin{aligned} \sum_{s=1}^z \{(r_{ps}-t)^2\} &= \sum_{i=1}^n \left(\sum_{\substack{1 \leq s \leq z \\ x_{is} > t}} w_i^2 (x_{is}-t)^2 \right) + \\ &\sum_{i=1}^n \left(w_i^2 \sum_{\substack{1 \leq s \leq z \\ x_{is} < t}} (x_{is}-t)^2 \right) + \\ &2 \sum_{1 \leq i < j \leq n} \left(w_i w_j \sum_{\substack{1 \leq s \leq z \\ x_{is} > t \\ x_{js} > t}} (x_{is}-t)(x_{js}-t) \right) + \\ &2 \sum_{1 \leq i < j \leq n} \left(w_i w_j \sum_{\substack{1 \leq s \leq z \\ x_{is} > t \\ x_{js} < t}} (x_{is}-t)(x_{js}-t) \right) + \end{aligned}$$

$$2 \sum_{1 \leq i < j \leq n} \left(w_i w_j \sum_{\substack{1 \leq s \leq z \\ x_{is} \leq t \\ x_{js} > t}} (x_{is} - t)(x_{js} - t) \right) +$$

$$2 \sum_{1 \leq i < j \leq n} \left(w_i w_j \sum_{\substack{1 \leq s \leq z \\ x_{is} < t \\ x_{js} < t}} (x_{is} - t)(x_{js} - t) \right).$$

Definition: The measure of risk Statistic Efficient, denoted by SE, can be defined as:

$$SE = \lambda M^- + \gamma M^+ + \sum_{1 \leq i < j \leq n} (\alpha_{ij} M_{ij}^{++} + \beta_{ij} M_{ij}^{+-} + \delta_{ij} M_{ij}^{-+} + \eta_{ij} M_{ij}^{--})$$

We can extend the sum of squared deviations and cross-squared deviations of returns from the target t to any order degree without loss of the properties displayed. One of the advantages of SE is that it provides a high dimensional space in the search region formed by the optimization problem and solved by using a QP algorithm that allows a global optimum. To compute the SE risk measure we can use a QP algorithm that, even though containing a higher number of parameters, requires much less computational consumption, compared to NSP models used on partial moments and on above and below a target risk measures.

3 A New Re-sampling model with Frequency Analysis

The EV model and its risk metrics evolution as MSTD model, by not taking into account the frequency distribution of the optimization results, do not allow calculating how frequently the optimal Portfolio occurs, so it is not possible to know how often the optimal Portfolio will be (was) optimal in future (past) periods. Among other reasons, due to the uncertainties in the estimates of position and dispersion measures of Assets, the Portfolio calculated by the EV model fails to attain the optimal parameters [7]. In addition, because Portfolio optimization models are based on the mean, variance and correlation

between Asset returns, persons often mistakenly assume normal distributions for the respective returns. The assumption of normality, however, is not a convenient assumption, the distributions of Assets returns are generally asymmetric and leptokurtic [8]. We did not find in the literature a Portfolio selection model that considers the frequency of occurrences as the model starting point and that considers as result the analysis of the frequency distribution of the optimal results, instead of statistical position and deviation measures. In the literature the Re-sampled Efficient Frontiers [12] [13] [14] model result is an qualitative average of re-sampled EV models, i. e., a statistical position measure. Also the Simulation-Optimization model [5], uses simulation only to create a disturbance around the input variables in order to obtain an output distribution where they can calculate a statistical position measure as an answer. The new re-sampling model with frequency analysis explained below use a new heuristic to calculate a probability measure of occurrence of the selected optimum. We will show through simulations that the consideration of the distribution frequencies of Ex-Post evaluation of the optimal parameters can provide the frequency of occurrence of the optimal Portfolio, that should be used on a Ex-Ante analysis. The heuristic for the new re-sampling model with frequency analysis, by performing the simulation, which considers all possible combinations of Assets and calculating from each iteration all possible optimal values, obtain the frequency distributions of the optimal parameters. Thus from the method of clusters, we find the centroid of the group with highest frequency of optimal results. The percentage invested in each Asset, the Asset return and the Asset risk are evaluated from the centroid of the group with highest frequency, which we defined as the Modal Optimum. We define the heuristic method to obtain the Modal Optimum as the Distribution Efficient Method (DEM). Thus, we obtain the Portfolio defined as Statistic and Distribution Efficient (SDE) from applying the DEM method to the SE risk measure. The centroid of the group with highest frequency is the representative of the Modal Optimum and the frequency of optimums in that group is the Modal Optimum frequency. By considering the Modal Opti-

imum from the frequency of occurrence of all the optimums selected for a group from a cluster analysis we had calculated the frequency of occurrence of an optimum by means of simulation where for each iteration we carried out an optimization. We consider the window of observations of Assets to be a unit window, where the expected return of the Portfolio is given by the return at each iteration. Thus, it is possible at each iteration to evaluate the optimal parameters, such as the percentage of capital invested in each Asset. We call the Portfolio return calculation with window for one period a naive Portfolio, since implicitly we assume that the current return is the best forecast for the average returns of the Assets. This hypothesis is similar to the assumption made for building the U-Theil statistic by [3], exhibited initially by [17]. We define the calculated optimal number of groups used in the cluster method by the V-Fold Cross Validation method [2]. In the process of exhibiting the results obtained by the SDE Portfolios, we use the optimal investment fraction defined by the DEM together with the holdout method. We use a data window for training, another data window for testing, and both for analyzing the performance of the SDE Portfolios. Specifically, we measure the SDE Portfolio performance, generated from the simulation, optimization and cluster analysis, by the Ex-Post and Ex-Ante analysis, applied to the data used to generate the Portfolio and to the holdout data, respectively. We will see two examples using the frequency analysis as a new form of the optimization result analysis and using the new concept of Modal Optimum. In that examples we compare the results obtained by the SE model to the results obtained from the EV model and from the MSTD model, with all them using the DEM to obtain the Modal Optimum.

4 Results

4.1 Ex-Post comparison of Portfolios with three Assets of the São Paulo Stock Exchange

In the Ex-Post analysis we compare the SDE Portfolio performance to the performance obtained from the EV and the MSTD Portfolios models (using the DEM to obtain the Modal Optimum) taking into account three Assets of the São Paulo Stock Exchange in Brazil (Bovespa), named PETR4, BBDC4 and GGBR4. This analysis considers a history Index window of 100 closing price (returns) of August 1, 2005 (100%) until December 22, 2005. The three Assets are among the most traded stocks of the Bovespa, usually preferred by the managers. Two of them have low risks (standard deviation) in comparison with other traded Assets and one has high risks.

We analyze the Ex-Post model frequency distribution of each model using the Modal Optimum by applying the DEM method to each model. The distribution of the three Bovespa Assets are obtained making the series stationary, with the finite difference of order one, and calculating the histogram of each Asset series, since these do not fit with goodness to any theoretical distribution usually fitted for Asset distribution, with asymmetric and leptokurtic returns [8]. The simulation of the three Bovespa Assets is performed by simulating the correlated histograms of each stationary Asset.

The comparison of the Portfolios models performance is done considering the EF of the EV model and setting the MRR by using the Capital Market Line (CML) as independently developed on the Capital Asset Pricing Model (CAPM), by [18][19], [16], [9] and [15]. We consider an MRR equals 4% for model comparison, after making the Assets stationary, in an EF varying from 0.5% to 6.0%, that means risk free Asset with return of near of 2.50% in the period of the window of 100 daily stationary Assets return; which is equivalent to risk free Asset with a return near of 110.00% per year. During the 100 daily returns the Selic tax (average adjusted rate of daily financing determined in the Special System of Settlement and Cus-

tody for federal securities by the Central Bank of Brazil) had an annual average of 119.14%.

In the Ex-Post frequency analysis, we applied the Modal Optimum fraction to invest in each Asset, determined by the compared Portfolio models, to measure the Portfolios returns generated by a simulation of 100,000 iterations, which allows the calculation of simulated variables with accuracy of 1% with 99% confidence. We re-sampled the historic window of 100 closing returns prices of the correlated histograms for the three Bovespa Assets.

To calculate the Modal Optimum and it frequency we used points on the efficient frontier with a MRR ranging from 0.5% to 6.0%. For being more detailed and comprehensive regarding the results, we show in Table 1 the frequency values of the Modal Optimum, for each considered model, on the range of the MRR. We can see in that example, for a specified MRR, that the frequencies of the Modal Optimum of the SDE model outperforms, with one exception, the results for the EV and the MSTD models.

Efficient Frontier MAR	MODAL OPTIMAL FREQUENCY		
	Markowitz	MSTD	SDE
0.50	77.49%	85.49%	99.45%
1.00	69.84%	70.64%	79.79%
2.00	58.04%	60.57%	67.86%
3.00	46.48%	50.96%	59.94%
4.00	30.13%	34.95%	48.70%
5.00	19.56%	22.92%	25.70%
6.00	25.07%	20.96%	22.61%
9.00	0.00%	0.00%	0.00%

Table 1: Modal Optimum Frequency

4.2 Ex-Ante comparison procedures for Portfolio Assets of the São Paulo Stock Exchange

We compare the SDE Portfolio performance to EV and the MSTD performance considering a Portfolio with Assets of the Bovespa, select by usual investor procedures. This analysis considers a history window of 100 closing price (returns) of August 1, 2005 until December 22, 2005 to calculate the Modal Optimum Portfolios and another sequential window of 395 returns, for holdouts, to be used on the Ex-Ante analysis, of December 23, 2005 until July 31, 2007. The Portfolio selection chose Assets from the Ibovespa (a theoretical Portfolio of Assets prepared in accordance with the criteria set out in a method of the Bovespa). The objective of the Ibovespa is to be the indicator of the average performance of the prices of most traded and representative Assets in the Brazilian stock market. The final Ex-Ante results also present a comparison with the Ibovespa Portfolio. The selected Assets are among the most traded Assets of the Bovespa, usually preferred by managers. We chose Assets with high liquidity and high weight in the theoretical Portfolio of the Ibovespa. We also chose Assets that will account for nearly 50% of the Ibovespa. This analysis considers two windows, one with 100 returns to calculate the Optimal Modal of each model, and another with 395 holdouts used in the Ex-Ante analysis. The range of the study comprises a high volatility (risk) market with two stress moments. The select Assets are PETR4, VALE5, BBDC4, USIM5, ITAU4, GGBR4, TNLP4, CSNA3, UBBR11, ITSA4, CMIG4 and BRKM5. After making the series stationary, during the first window of 100 returns used for the Portfolios models calculus, we classified the selected Assets by their annual standard deviation as low volatility and high volatility Assets, as shown by the standard deviation of Assets, presented in Table 2.

In addition to the usual restrictions to the EV model, we also considered: i) not using the Asset TNLP4, which presents problems with minority shareholders; ii) that the percentage of an Asset with low volatility should always be less than 35%; and iii) that the percent-

Standard Deviation	Low Volatility							High Volatility				Not Used
	BRKM5	CMIG4	VALE5	ITAU4	PETR4	ITSA4	CSNA3	GGBR4	UBBR11	BBDC4	USIM5	TNLP4
	2.19	2.22	2.24	2.40	2.44	2.45	2.46	2.75	2.91	3.00	3.10	3.46

Table 2: Assets Volatilily

age of an Asset with high volatility should always be less than 25%.

For evaluate the Modal Optimum we generated a simulation of 100,000 iterations, which allows the calculation of simulated variables with accuracy of 1% with 99% confidence. We re-sampled the historic distribution of the window of 100 closing prices of the correlated Asset returns for the select Assets from the Bovespa. The distribution of the eleven Bovespa Assets is obtained making the series stationary and calculating the histogram of each series, the simulation of the eleven Bovespa Assets is performed by simulating the histograms of each Asset, considering the Assets correlation. For models comparison we have the MRR equals 3,0%, by the CML in a EF varying from 0.5% to 6.0%, that means a risk free Asset of 2.50% in the period of the window of 100 daily stationary returns. In the Ex-Ante analysis, using the frequency Ex-Post analysis, we applied the Modal Optimum fraction to invest in each Asset, determined by the compared Portfolio models and evaluated using the first window of 100 returns, to the second window of 395 holdouts returns. The investor uses the Modal Optimum Portfolio obtained by the DEM method, which will be the Portfolio of Assets acquired for the next period and reinvested in each new period as in Markowitz[13].

Making the beginning of the holdout period equals 100%, Figure 1 shows the results for applying the Modal Optimum of each model to the second window of 395 holdouts returns, the graph also include the Ibovespa holdout returns. Table 3 shows the index Return results for the EV and the MSTD models using the DEM method, the SDE models and for the Ibovespa at the end of the 395 holdout window, making the beginning of the holdout period equals 100%.

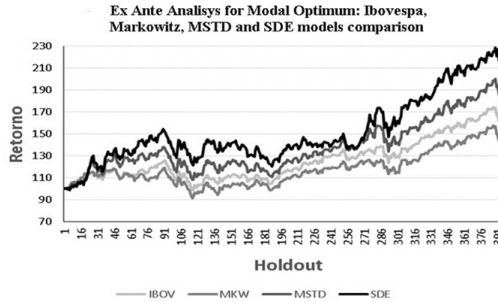


Figure 1: Ex-Ante Analysis

Ibovespa	Markowitz	MSTD	SDE
162.56%	145.91%	187.34%	221.79%

Table 3: Results at the End of the Holdout Window

5 Conclusion

The evolution of the E-V Efficient Portfolio [11] for models based on risk measures that use partial moments, downside risk, upside risk [1] and [4], LPM [6], LTD and ATD [8] does not provide a solution to the optimization problem with a literal expression, but transforms the optimization algorithms into a NSP. These kinds of models have a complex solution with high computational consumption and provide only a single viable solution or a local optimum.

We presented a new measure of risk that considers all cross interrelationships between returns in addition to deviations above and below a reference target. Our measure provides a high dimensional space in the search region formed by the optimization problem and solved by the optimization algorithm. This allows us to try obtaining solutions with higher returns compared to the EV and MSTD models. We also have as solution of the optimization: a QP problem with a optimal

solution that presents lower computational consumption compared to NSP models. Moreover, we used simulations and a new heuristic that add to the usual optimization procedures by re-sampling the Portfolio in order to consider the Ex- Post returns distributions in addition to the position and dispersion measurements commonly used. This allowed us to evaluate uncertainties inherent to the process of Portfolio selection and access the Modal Optimum, giving a probability measure to the occurrence of the selected optimal, used on the ex-Ante Analysis. Our paper makes contributions to the development of the Portfolio Optimization Models shedding some light on the use of Modal Optimums and introducing Frequency Analyses in Portfolio Optimization.

References

- [1] Bawa, V. S. (1975). Optimal rules for ordering uncertain prospects. *Journal of Financial Economics*; Volume 2, Issue 1, March 1975, pp. 95 -121.
- [2] Burman P. (1989). A Comparative Study of Ordinary Cross-Validation, V-Fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika* Vol. 76, No. 3 (Sep., 1989), pp. 503-514.
- [3] Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting*, 8, 81-98.
- [4] Fishburn, P.C. (1977). Mean-Risk Analysis with Risk Associated with Below-Target Returns. *The American Economic Review* Vol. 67, No. 2 (Mar., 1977), pp. 116-126, Published by: American Economic Association.
- [5] Glover et al. (2003). Practical introduction to simulation optimization. *Simulation Conference, 2003, Proceedings of the 2003 Winter*, vol.1, pp.71,78 Vol.1, 7-10 Dec. 2003.
- [6] Holthausen, D. M, (1981). A Risk-Return Model with Risk and Return Measured as Deviations from a Target Return. *American Economic Review*, American Economic Association, vol. 71(1), pp. 182-88, March.

- [7] JJobson, J.D.; and Korkie, B. (1981). Putting Markowitz Theory to Work. *The Journal of Portfolio Management*, Summer 1981, Vol. 7, No. 4: pp. 70-74.
- [8] Kang et al. (1996). The mean-separated target deviations risk model. *Journal of Economics and Business*; 48:47-66. Temple University.
- [9] Lintner, John (1965a,b).The valuation of risk assets and the selection of risky investments in stock Portfolios and capital budgets, *Review of Economics and Statistics*, 47 (1), pp. 13 - 37.
- [10] Markowitz H.M. (1952). Portfolio Selection. *The Journal of Finance*, Vol. 7, No. 1. (Mar., 1952), pp. 77-91.
- [11] Markowitz H.M. (1991) Foundations of Portfolio Theory. *The Journal of Finance* Vol. 46, No. 2 (Jun, 1991), pp. 469-477.
- [12] Michaud, R. O. (1998). *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*. Boston, MA: Harvard Business School Press, 1998.
- [13] Michaud, R. O. (2003). An Examination of Re-sampled Portfolio Efficiency: A Comment by Michaud, R. O. *Financial Analysts Journal*, January/February 2003, Vol. 59, No. 1: 15-16.
- [14] Michaud, R. O. (2013). Deconstructing Black - Litterman: How to get the Portfolio you already knew you wanted. *Journal of Investment Management*, Vol. 11, No. 1, (2013), pp. 6 - 20; JOIM 2013.
- [15] Mossin, Jan. (1966). Equilibrium in a Capital Asset Market, *Econometrica*, Vol. 34, No. 4, pp. 768 - 783.
- [16] Sharpe, William F. (1964).Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance*, 19 (3), pp. 425 - 442
- [17] Theil, H. (1966). *Applied Economic Forecasting*. North-Holland Publ. Co. Amsterdam, 1966.
- [18] Treynor, Jack L. (1961). Market Value, Time, and Risk. Unpublished manuscript.
- [19] Treynor, Jack L. (1962).Toward a Theory of Market Value of Risky Assets. Unpublished manuscript. A final version was published in 1999, in *Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics*. Robert A. Korajczyk (editor) London: Risk Books, pp.15 - 22.

Modelling (and forecasting) extremes in time series: A naive approach


M. Manuela Neves

Instituto Superior de Agronomia, and CEAUL, Universidade de Lisboa, Lisboa, Portugal, manela@isa.ulisboa.pt

Clara Cordeiro


Faculdade de Ciências e Tecnologia, Universidade do Algarve, and CEAUL, Universidade de Lisboa, Lisboa, Portugal, ccordei@ualg.pt

Keywords: Extreme Value Theory; Extremal index estimation; Resampling procedures; Time Series.

Abstract: In *Extreme Value Theory*, we are essentially interested in the estimation of quantities related to extreme events. Whenever the focus is in large values, estimation is usually performed based on the largest k order statistics in the sample or on the excesses over a high level u . Here we are interested in modelling (and forecasting) extremes in time series. For modelling and forecasting classical time series, Boot.EXPOS is a computational procedure built in the  environment that has revealed to perform quite well in a large number of forecasting competitions. However, to deal with extreme values, a modification of that algorithm needs to be considered and is here under study.

1 Introduction and Motivation

Time series analysis deals with records that are collected over time. The records are usually dependent, and the time order of data is important. Depending on the application, data may be collected hourly, daily, weekly, monthly, yearly, etc. Time series arise in many different contexts. Its impact on scientific, economic and social applications is well recognized by the large list of fields in which impor-

tant time series problems may arise. Time series can show different displays. Let us illustrate a few time series, two of them existing in the  packages `datasets` and `fma`, see Fig.1.

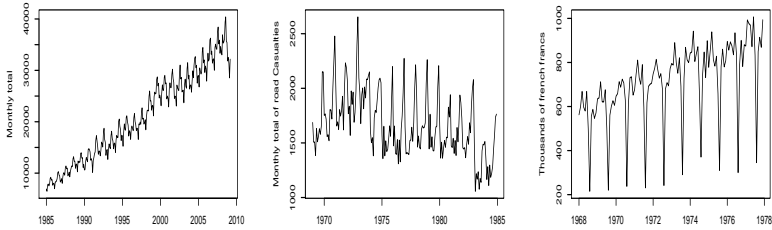


Figure 1: Number of airplanes in the FIR Lisbon(left), see [6]; Deaths and serious injuries on UK roads (center); and Sales of printing and writing paper (right).

In time series analysis, there are several challenging topics among which the treatment of extreme values has been capturing the interest of researchers. Modelling and predicting the behaviour of extreme (often maximum) values of the time series (e.g. security reasons) need special procedures.

The paper is structured as follows. In Section 2, basic results in extreme value theory both for independent and for dependent sequences are briefly reviewed. A new parameter that can appear in the limit law of the maximum of a stationary sequence, under some conditions, is described. Resampling techniques and their application together with exponential smoothing methods for modelling and prediction of a time series are reviewed in Section 3. In this section, a modification of that computational procedure, already introduced in [27] is again considered and used in extreme value theory estimation. More efficient bootstrap procedures can lead to more reliable estimates.

2 Basics in statistical analysis of univariate extremes

Statistical analysis of the extremes in time series was initially dedicated to problems in hydrology and insurance, but in the last decades the applications have spread out to a huge variety of areas, such as climatology, finance, environmental sciences (here mainly because of the direct impact in the society), etc.

The classical limiting results in Extreme Value Theory (EVT) were initially obtained through arguments that assumed an underlying process consisting of a sequence of independent and identically (i.i.d.) random variables, (X_1, \dots, X_n) , with common and unknown distribution F . Suppose we want to know the distribution of $M_n \equiv X_{n:n} := \max(X_1, \dots, X_n)$.

Given that $X_{n:n} \xrightarrow{\mathbb{P}} x_F =: \sup\{x \in \mathbb{R} : F(x) < 1\}$, the right end-point of F , we are facing the situation of a degenerate distribution. First results for the existence of a non-degenerate limit for that probability date back to the beginning of the last century but were completely established by [12] and [16] that gave conditions for the existence of sequences $\{a_n\} \in \mathbb{R}^+$ and $\{b_n\} \in \mathbb{R}$ such that,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{M_n - b_n}{a_n} \leq x \right) = \lim_{n \rightarrow \infty} F^n(a_n x + b_n) = \text{EV}_\xi(x), \quad (1)$$

when $n \rightarrow \infty$ and $\forall x \in \mathbb{R}$. EV_ξ is a nondegenerate distribution function. It is called *Extreme Value* d.f., and is given by

$$\text{EV}_\xi(x) = \begin{cases} \exp[-(1 + \xi x)^{-1/\xi}], & 1 + \xi x > 0 & \text{if } \xi \neq 0 \\ \exp[-\exp(-x)], & x \in \mathbb{R} & \text{if } \xi = 0, \end{cases}$$

where ξ , the extreme value index, is the primary parameter in extreme value theory because it measures the weight of the right tail function, $\bar{F} = 1 - F$, of the underlying model.

A function F for which the limit in (1) holds is said to be in the max-domain of attraction of EV_ξ , and we write $F \in \mathcal{D}_\mathcal{M}(\text{EV}_\xi)$.

These models can also incorporate location (λ) and scale ($\delta > 0$) parameters, and are generally represented by

$$\text{EV}_\xi(x; \lambda, \delta) \equiv \text{EV}_\xi((x - \lambda)/\delta).$$

2.1 From i.i.d. to a dependent set-up

In many applications, temporal independence is unrealistic. Whenever the original scheme is no longer identically distributed, but it remains independent, those limiting results may hold true. However, when it is not possible to assume independence, we are faced with new situations. For many real problems the stationarity is the first realistic situation to be considered. In the last decades, many progresses have been made in parameter estimation of extreme values in time series, with relevance to asymptotic results. By the 1990s there was an increased interest in extremal time series, see [29, 5, 2, 4, 24], to mention a few.

Temporal dependence is common in univariate extremes of time series leading to clusters of extremes, which means that extreme values are likely to occur in temporal proximity. An excellent overview of the topic of extremal clustering is provided by [8].

As an illustration, let us consider the following sequences:

Example 2.1 *Let $\{X_n\}$ be a sequence of i.i.d. variables from the model $F(x) = (1 - \exp(-x))^2$, $x \geq 0$, and $\{Y_n\}_{n \geq 1}$ a two-dependent sequence defined by $Y_n = \max(Z_{n+1}, Z_n)$, $n \geq 1$, where Z_n are unit exponential i.i.d..*

We have then the underlying model for Y_n given by $F(y) = \mathbb{P}[Z_{n+1} \leq y, Z_n \leq y] = (1 - \exp(-y))^2$ $y \geq 0$.

Plotting some values from $\{X_n\}$ and from $\{Y_n\}$, clusters of exceedances of high levels of size equal to 2, for the $\{Y_n\}$ sequence, can be seen, Fig.2. It can also be seen a *shrinkage* of the largest observations for the 2-dependent sequence, although we have the same model underlying both sequences.

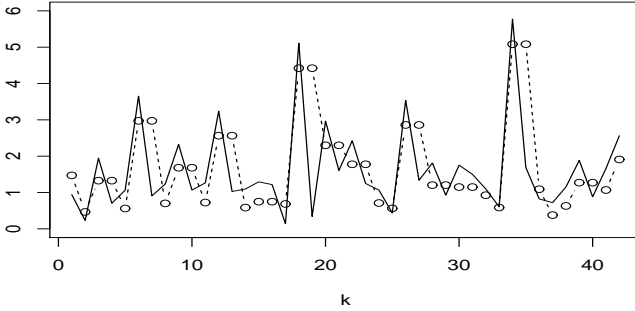


Figure 2: One realization of an i.i.d. process (solid) and a 2-dependent process (dot dash) with the same marginal d.f.

Let $\{X_n\}_{n \geq 1}$ be a stationary sequence. Under adequate conditions, the d.f. of the maximum, $X_{n:n}$, of a stationary sequence may be directly related to the maximum $Y_{n:n}$ of the associated i.i.d. sequence, through a new parameter, the so-called *extremal index*. The extremal index, θ , $0 < \theta \leq 1$, appears as

$$\mathbb{P}(X_{n:n} \leq x) \approx F^{n\theta}(x) \approx \text{EV}_\xi \left(\frac{x - b'_n}{a'_n} \right) \quad \left\{ \begin{array}{l} a'_n = a_n \theta^\xi \\ b'_n = b_n + a_n \frac{\theta^\xi - 1}{\xi} \end{array} \right.$$

In [23] conditions were established under which a stationary sequence has the same limiting EV_ξ as the associated i.i.d. sequence, but different scale and location parameters,

$$\lambda_\theta = \lambda + \delta \frac{\theta^\xi - 1}{\xi}, \quad \delta_\theta = \delta \theta^\xi \quad \xi_\theta = \xi,$$

where (λ, δ, ξ) are the location, scale and shape parameters of EV_ξ , respectively. A reliable estimation of θ is then required, not only by itself but because of its influence on the estimation of other parameters of interest.

2.2 The extremal index and its estimation

One common interpretation of θ is as being the reciprocal of the “mean time of duration of extreme events” which is directly related to the exceedances of high levels, see [20, 22]. Parameter θ can then be defined as $\theta = 1/(\text{limiting mean size of clusters})$.

Now, identifying clusters by the occurrence of downcrossings (or upcrossings), we can write

$$\theta = \lim_{n \rightarrow \infty} \mathbb{P}[X_2 \leq u_n | X_1 > u_n] = \lim_{n \rightarrow \infty} \mathbb{P}[X_1 \leq u_n | X_2 > u_n]$$

and the interpretation of θ has suggested the so-called Up-Crossing estimator, see [25, 10, 11], defined as:

$$\hat{\Theta}_n^{UC} := \frac{\sum_{i=1}^{n-1} I(X_i \leq u_n < X_{i+1})}{\sum_{i=1}^n I(X_i > u_n)}, \quad (2)$$

where $I(A)$ is the indicator function of A . Consistency of this estimator is obtained provided that the high level u_n is a normalized level, i.e. if with $\tau \equiv \tau_n$ fixed, the underlying distribution function F verifies

$$F(u_n) = 1 - \tau/n + o(1/n), \quad n \rightarrow \infty \quad \text{and} \quad \tau/n \rightarrow 0.$$

Other estimators have appeared in the literature, motivated by other forms of cluster identification, such as the blocks estimator and the runs estimator, see [18, 19, 30, 31]. Conditions for the asymptotic normality of those estimators can be seen in [19, 30, 31].

As usual in semiparametric context, the estimators considered, despite having good asymptotic properties, present high variance for high levels *vs* high bias when the level decreases, showing then a strong dependence on the high threshold u_n , for finite samples.

3 Resampling procedures

Resampling computer-intensive methodologies, like the generalised jackknife, [15], and the bootstrap, [9], have been revealing them-

selves as important tools for a reliable semi-parametric estimation of parameters of extreme events. Let us briefly see the application of those methodologies in the θ estimation.

3.1 The Generalized Jackknife methodology

By using generalized jackknife methodology, [13] proposed a reduced-bias Generalized Jackknife estimator of order 2, $\hat{\Theta}^{GJ}$, based on the estimator $\hat{\Theta}^{UC}$ computed at three levels: k , $\lfloor k/2 \rfloor + 1$ and $\lfloor k/4 \rfloor + 1$, ($\lfloor x \rfloor$ – integer part of x), defined as

$$\hat{\Theta}^{GJ}(k) := 5\hat{\Theta}^{UC}(\lfloor k/2 \rfloor + 1) - 2(\hat{\Theta}^{UC}(\lfloor k/4 \rfloor + 1) + \hat{\Theta}^{UC}(k)). \quad (3)$$

More generally [13] considered the levels k , $\lfloor \delta k \rfloor + 1$ and $\lfloor \delta^2 k \rfloor + 1$, depending on the *tuning parameter* δ , $0 < \delta < 1$, and got then a class of estimators. Actually $\hat{\Theta}^{GJ}$, in (3), is obtained with $\delta = 1/2$. This estimator illustrates the simulation study performed with the “Max-Autoregressive Process (ARMAX process)”, see [2].

Example 3.1 Let $\{Z_i\}_{i \geq 1}$ be a sequence of independent, unit-Fréchet distributed random variables. For $0 < \theta \leq 1$, let

$$X_1 = Z_1 \quad X_i = \max\{(1 - \theta)X_{i-1}, \theta Z_i\} \quad i \geq 2.$$

For $u_n = nx$, $0 < x < \infty$, $\mathbb{P}\{M_n \leq u_n\} \rightarrow \exp(-\theta/x)$, as $n \rightarrow \infty$, being θ the extremal index of the sequence.

The reduced-bias estimator in (3) outperforms the associated classical estimator. However, for a given sample, the choice of the number of upper order statistics to be used is a difficulty not yet solved. See Figure 3, where three different samples were generated, considering three different values for the parameter, in an ARMAX model. The estimates paths show how difficult it is to choose k and to obtain a reliable estimate of θ .

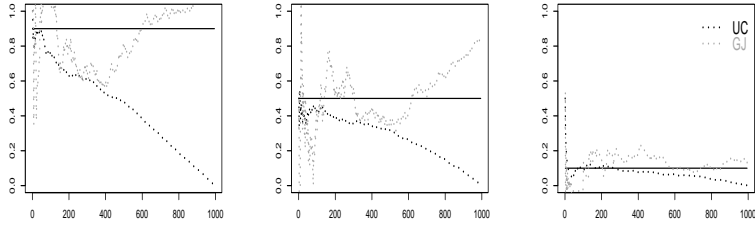




Figure 3: One sample path for UC and GJ estimates in the ARMAX model for three simulated samples with $\theta = 0.9, 0.5, 0.1$ (from the left to the right).

3.2 The bootstrap under dependence

For modelling and forecasting time series [6, 7] developed a computational procedure, built in the  environment, based on Exponential Smoothing Methods jointly with “adequate” bootstrap procedures. When applied to a large set of time series, competitive results were obtained compared with the best procedures available, see [7].

So the main motivation of this work is to explore and to modify that automatic procedure in order it can be an alternative for modelling and (forecasting) extreme values in time series. Preliminary results have been presented in [27] and are used here in the θ estimation.

The aforementioned computational procedure, for modelling and forecasting time series, chooses among a set of models, that one that best fits the data. Sieve bootstrap principle is applied to the residuals; an autoregressive model with increasing order is fitted to the residuals; stationarity is tested; transformations or differentiations are performed when necessary, and after bootstrapping the second level of residuals a bootstrap estimated series is obtained. Forecast is performed based on the bootstrap estimated values and on the model parameters estimated at the initial step. Measures of forecast errors are also included in the algorithm. A description and sketch of the algorithm is presented in [6, 7, 26] among others.

Fig.4 illustrates the result of forecasting twelve months applying Boot.EXPOS and *ets*⁷ [21], using the dataset *UKDriverDeaths* available in . The good

⁷Stands for error, trend and seasonality.

performance of the Boot.EXPOS procedure is clearly illustrated both for point forecast values and for forecasting intervals.

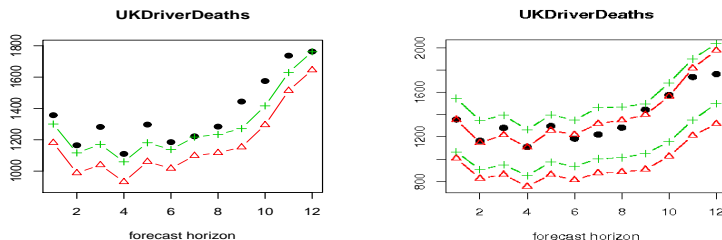


Figure 4: True values (●) compared with Boot.EXPOS values (– + –) and ets values (– △ –).

3.3 Modelling time series extremes

The classical bootstrap does not work in a dependent context. This was referred to [3] and later in [1], who showed that in extreme value theory the bootstrap version for the maximum (or minimum) does not converge to the extremal limit laws. Actually, [32] pointed out “... to resample the data for approximating the distribution of the k largest observations would not work because the “pseudo-samples” would never have values greater than $X_{n:n}$ ”. A bootstrap method considering to resample a smaller size than the original sample was proposed in [17] for estimating mean squared error and smoothing parameter in nonparametric problems. The idea in [17] was to choose the resample size, n_1 , to be less than the original sample size, n , and use knowledge of the amount by which the two samples differ to estimate mean squared error and to select the optimal smoothing parameter for deriving a bootstrap estimator of a functional of (X_1, \dots, X_n) . He suggested resampling a subsample of size $n_1 = O(n^{1-\epsilon})$ with $0 < \epsilon < 1$. The procedure developed in [17] was illustrated for nonparametric density estimation, nonparametric regression and tail parameter estimation. In this latter case, the tail parameter estimators in a semi-parametric approach need an adequate choice of the number, k , of upper order statistics, that should be chosen such that the asymptotic mean squared error of the estimator is minimized. The [17] bootstrap procedure suggests the following: to draw a resample of size n_1 from the original sample of size n , to obtain the bootstrap estimate of the mean squared error of the estimator

considered, let us denote it as $\widehat{\text{MSE}}(n_1, k_1)$, where k_1 are the upper order statistics of the n_1 -sized resample. Supposing that the asymptotically optimal k is of the form Cn^γ , where $0 < \gamma < 1$ is a known constant and C is unknown, what is a common result, [17] proposed, for a given class of models, if the optimal k_1 is asymptotic to Cn_1^γ , then

$$\hat{k}_0 \simeq \hat{k}_{1,0}(n/n_1)^\gamma, \quad (4)$$

is asymptotic to Cn^γ . For several models, [17] showed that $\gamma = 2/3$. This idea was exploited in a very preliminary study in [27], where the functional under study was the *maximum*, taking advantage of the good performance of *Boot.EXPOS* for modelling and forecasting time series. A subsample of size $n_1 = \lfloor n^{0.995} \rfloor$ of the residuals in the algorithm was considered. Values of the resampled series were then “improved” on basis of the relation (4) – this is now called *Boot.EXPOS with subsampling*. See Fig.5 as an illustration.

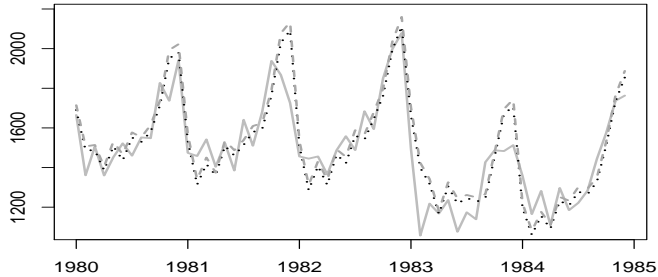


Figure 5: Subset of observed UKDriverDeaths values (solid grey) and forecasts obtained using *Boot.EXPOS* (dashed), *Boot.EXPOS with subsampling* (dotted).

The *Boot.EXPOS with subsampling*, was applied to a simulated data set and to a real data set, the *UKDriverDeaths* time series. The interest is to estimate θ . Figures 6 and 7 show sample paths for the θ -estimator, $\hat{\Theta}_n^{UC}$, in (2), and $\hat{\Theta}^{GJ}$, in (3), and the associated bootstrap estimates calculated using *Boot.EXPOS with subsampling*, $\hat{\Theta}_n^{UC*}$ and $\hat{\Theta}^{GJ*}$, respectively.

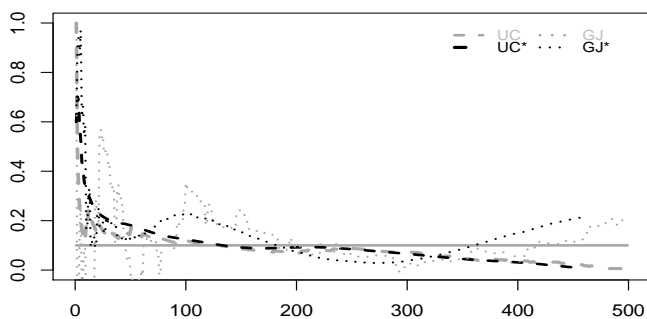


Figure 6: UC and GJ θ -estimates in an ARMAX process with $\theta = 0.1$. UC* and GJ* θ -estimates using Boot.EXPOS with subsampling.

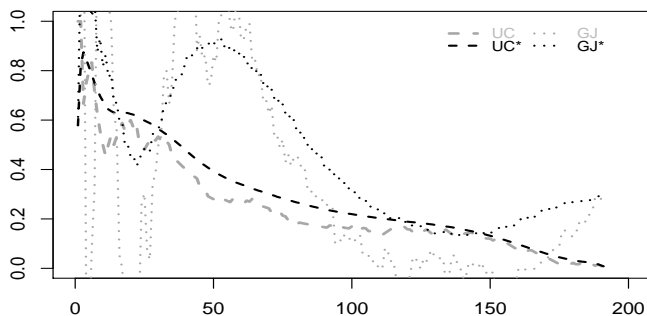


Figure 7: UC and GJ θ -estimates in the UKDriverDeaths time series and the associated UC* and GJ* θ -estimates using Boot.EXPOS with subsampling.

4 A brief discussion

The procedure here proposed and based on [17] results seems to be a promising bootstrap approach for modelling and forecasting extremes,

providing more stable paths to the parameters estimates. Other values for the θ parameter in the ARMAX process have been considered, leading to similar results, not shown for reasons of space. More research needs to be performed. A large simulation study is now in progress.

Acknowledgements

We would like to thank the two anonymous referees for their valuable and helpful comments, that highly improve this version of the manuscript. Research partially supported by National Funds through Fundação para a Ciência e a Tecnologia, Portugal, through the project FCT Portugal UID/MAT/00006/2019. The authors' thanks to **Portugal Navigation-NAV Portugal, E.P.E.** for providing the data.

References

- [1] Angus, J. (1993). Asymptotic theory for bootstrapping the extremes. *Communication in Statistics Theory and Methodology*, 22(1), 15–30.
- [2] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J.L. (2004). *Statistics of Extremes. Theory and Applications*. England, John Wiley & Sons.
- [3] Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9, 1196–1217.
- [4] Chavez-Demoulin, V. and Davison, A.C. (2012). Modelling Time Series Extremes. *Revstat-Statistical Journal*, 10:1, 109–133.
- [5] Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, New York.
- [6] Cordeiro C. and Neves M. (2009). Forecasting time series with Boot.EXPOS procedure. *Revstat*, 7:2, 135–149.
- [7] Cordeiro, C. and Neves M.(2010). Boot.EXPOS in NNGC competition): *Proceedings of the IEEE World Congress on Computational Intelligence (WCCI 2010)*, 1135–1141.
- [8] Davis, R.A., Mikosch, T. and Zhao, Y. (2013). Measures of serial extremal dependence and their estimation. *Stochastic Processes and Their Applications* 123: 7, 2575–2602

- [9] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 7:1–26.
- [10] Gomes, M.I. (1990). Statistical inference in an extremal markovian model. *COMPSTAT*, 257–262.
- [11] Gomes, M.I. (1992). Modelos extremais em esquemas de dependência. *I Congresso Ibero-Americano de Estadística e Investigación Operativa*, 209–220.
- [12] Gnedenko, B.V. (1943). Sur la distribution limite d’une série aléatoire. *Annals of Mathematics* 44, 423–453.
- [13] Gomes, M.I., Hall, A. and Miranda, C. (2008). Subsampling techniques and the Jackknife methodology in the estimation of the extremal index. *Computational Statistics and Data Analysis* 52:4, 2022–2041.
- [14] Gomes, M.I., Figueiredo, F. and Neves, M.M. (2011). Adaptive estimation of heavy right tails: the bootstrap methodology in action. *Extremes*, DOI: 10.1007/s10687-011-0146-6
- [15] Gray, H.L., and Schucany, W.R. (1972). *The Generalized Jackknife Statistic*. Marcel Dekker.
- [16] de Haan, L. (1970). *On regular variation and its application to the weak convergence of sample extremes*, Amsterdam, Mathematisch Centrum.
- [17] Hall P. (1970). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate Analysis*, 32, 177–203.
- [18] Hsing, T. (1991). Estimating the parameters of rare events. *Stochastic Processes and Their Applications*, 37:117–139.
- [19] Hsing, T. (1993). Extremal index estimation for a weakly dependent stationary sequence. *The Annals of Statistics*, 21:2043–2071.
- [20] Hsing, J. T., Hüsler, J. and Leadbetter, M. R. (1988). On the exceedance point process for a stationary sequence. *Probability Theory and Related Fields*, 78, 1: 97–112.
- [21] Hyndman, R., A. Koehler, J. Ord and R. Snyder (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer-Verlag.

- [22] Leadbetter, M.R. and Nandagopalan, L. (1989). On exceedance point process for stationary sequences under mild oscillation restrictions. In *Extreme Value Theory: Proceedings, Oberwolfach 1987*, J. Hüsler and R.D. Reiss (eds.), Lecture Notes in Statistics 52, 69–80. Springer-Verlag, Berlin.
- [23] Leadbetter, M., Lindgren, G. and Rootzen, H. (1983). *Extremes and related properties of random sequences and series*. Springer-Verlag, New York.
- [24] McElroy, T. (2016). On the measurement and treatment of extremes in time series. *Extremes*, 16:3, 467–490
- [25] Nandagopalan, S. (1990). *Multivariate Extremes and Estimation of the Extremal Index*. PhD Thesis, University of North Carolina, Chapel Hill.
- [26] Neves, M. and Cordeiro, C. (2011). Exponential smoothing and re-sampling techniques in time series prediction. *Discussiones Mathematicae, Probability and Statistics*, 30, 87–101.
- [27] Neves, M. M. and Cordeiro, C. (2014). *Statistical modelling in time series extremes: an overview and new steps*. In Gilli, M., Gonzalez-Rodriguez, G. and Nieto-Reyes, A. (eds.).): *Proceedings of COMP-STAT 2014*.
- [28] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- [29] Reiss, R-Dieter, Thomas, M. (1997). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhauser.
- [30] Smith, R. and Weissman, I. (1994). Estimating the extremal index. *Journal of the Royal Statistical Society B* 56:515–528.
- [31] Weissman, I. and Novak, S. (1998). On blocks and runs estimators of the extremal index. *Journal of Statistical Planning and Inference* 66:281–288.
- [32] Zelterman D.(1993, 2012). A Semiparametric Bootstrap Technique for Simulating Extreme Order Statistics, *Journal of the American Statistical Association*, 477–485.

A importância dos conceitos e das classificações nas Estatísticas da Educação

Nuno Rodrigues

Direção-Geral de Estatísticas da Educação e Ciência (DGEEC),

nuno.rodrigues@dgeec.mec.pt

Joaquim Santos

DGEEC, *joaquim.santos@dgeec.mec.pt*

Carlos Malaca

DGEEC, *carlos.malaca@dgeec.mec.pt*

Luísa Canto e Castro Loura

DGEEC e CEAUL, *luisa.castro@dgeec.mec.pt*

Palavras-chave: Estatísticas da educação; Conceitos de produção e difusão estatística; Classificação CITE /ISCED.

Resumo: Este artigo tem por objetivo apresentar as especificidades de dois importantes instrumentos de suporte à preparação das estatísticas oficiais nas áreas da Educação e Formação - os conceitos de produção e difusão estatística e a Classificação Internacional Tipo da Educação (CITE) 2011 - e dar alguns exemplos práticos dos impactos que diferentes opções, quer no que respeita aos conceitos, quer no que respeita à CITE, podem ter nos indicadores oficiais de educação e formação.

1 Introdução

A necessidade de dispor de informação estatística nos diferentes domínios - entre os quais a Educação - tem vindo a ser progressivamente compreendida a nível global. A esta evolução na perceção

social da importância das estatísticas não se pode dissociar o esforço conjunto de todos os intervenientes, quer sejam as autoridades estatísticas responsáveis pela recolha, quer sejam as entidades ou indivíduos responsáveis pelo reporte da informação.

Para se compreender o papel das diferentes autoridades estatísticas (tanto a nível nacional, como europeu), importa partilhar:

- a **visão do Sistema Estatístico Europeu**, que “... com base em princípios e métodos científicos, o Sistema Estatístico Europeu proporcionará e melhorará continuamente um programa de estatísticas europeias harmonizadas que constituirá uma base essencial dos processos democráticos e dos progressos sociais”
- a **respetiva missão**, de prestar “(...) à União Europeia, ao mundo e ao público informação independente de grande qualidade sobre a economia e a sociedade, a nível europeu, nacional e regional, e disponibilizar publicamente essa informação, para efeitos de apoio ao processo de decisão, de investigação e de debate” [1].

Para garantir a qualidade das estatísticas e a possibilidade de comparação internacional, torna-se imperioso desenvolver e adaptar os diversos mecanismos de recolha e produção de estatísticas a terminologias e classificações comuns. É também importante garantir a existência de dados harmonizados. Esse objetivo alcança-se, através da utilização de dois instrumentos fundamentais: os conceitos de produção e difusão estatística nas áreas da Educação e Formação e a Classificação Internacional Tipo da Educação (CITE).

Neste artigo pretende-se, para além de uma breve abordagem ao Sistema Estatístico Nacional, apresentar com maior detalhe as especificidades dos dois instrumentos anteriormente referidos e dar alguns exemplos práticos dos impactos que diferentes opções, quer no que respeita aos conceitos, quer no que respeita à CITE, podem ter nos indicadores oficiais de educação e formação.

2 O Sistema Estatístico Nacional

O Sistema Estatístico Nacional (SEN) define as normas e toda a estrutura que deve orientar a produção de estatísticas oficiais em Portugal [2]. Definindo direitos e deveres das autoridades estatísticas, das pessoas (individuais ou coletivas) que fornecem os dados e dos utilizadores, o SEN rege-se por 6 princípios principais: 1) Autoridade estatística (que define os aspetos relacionados com os processos de recolha de dados); 2) Independência técnica (das autoridades e dos seus colaboradores); 3) Segredo estatístico (de forma a aumentar a confiança no sistema, salvaguardando a privacidade de todos os cidadãos e entidades que reportem informação); 4) Qualidade (cumprindo os padrões nacionais e internacionais); 5) Acessibilidade estatística (cumprir calendários de divulgação e ir ao encontro das necessidades dos utilizadores); 6) Cooperação entre autoridades estatísticas.

O SEN integra o Conselho Superior de Estatística (CSE), órgão que orienta e coordena o Sistema, o Instituto Nacional de Estatística, I. P. (INE), o Banco de Portugal, os Serviços Regionais de Estatística das Regiões Autónomas dos Açores e da Madeira e as entidades produtoras de estatísticas oficiais por delegação do INE (entre as quais a DGEEC - Direção-Geral de Estatísticas da Educação e Ciência). Como definido no Regulamento Interno, o CSE órgão funciona em plenário, sessões restritas e secções permanentes ou eventuais, podendo estas decidirem pela criação de grupos de trabalho com o intuito de estudar determinadas matérias mais específicas.

Foi com este enquadramento legal que, em 2010, a Secção Permanente de Estatísticas Sociais, do Conselho Superior de Estatística, aprovou a constituição do Grupo de Trabalho de Estatísticas da Educação e Formação (GTEEF). O GTEEF, que integrava, entre outras entidades, o INE e a DGEEC, tinha como mandato 9 linhas principais, das quais se destacam o acompanhamento e promoção da atualização dos conceitos para fins estatísticos nas áreas da “educação e formação” e o acompanhamento da implementação da nova *International Standard Classification of Education* (ISCED) e a sua

tradução e adaptação para a língua portuguesa e para o sistema educativo nacional.

3 Produção e difusão estatística nas áreas da Educação e Formação

A correta definição de conceitos, com a atualização e transversalidade necessária, é fundamental para a correta definição metodológica de indicadores estatísticos e consequente interpretação dos resultados obtidos. Assim, e considerando que há praticamente 10 anos que não se procedia a alterações, a revisão e atualização do conjunto de conceitos utilizados nas áreas da educação e formação procurou que a análise e a validação efetuada se ajustassem à realidade atual - e previsivelmente futura.

3.1 Metodologia e constrangimentos

Em termos metodológicos, o trabalho realizado no âmbito do GTEEF passou pela sistematização dos conceitos existentes no Sistema de Metainformação do INE e pela apresentação, exposição e utilização de critérios subjacentes à metodologia de trabalho [3] [4].

Não esquecendo a necessidade de definir um conjunto de conceitos rigoroso, consistente, coerente e sistematizado, foi necessário: 1) optar por definições latas, abrangendo a realidade de diferentes produtores e utilizadores de informação e evitando conceitos de conteúdo similar; 2) evitar a “multiplicação desnecessária de conceitos”; 3) e torná-los de aplicação transversal (i.e., por haver sido definido, aclarado, discutido e acordado por todas as entidades, reúne as condições necessárias para ser “universalmente” reconhecido e adotado).

Ao longo do trabalho, foram analisados 615 conceitos. Entre os que tiveram maior discussão, encontram-se os conceitos de aluno, ano de escolaridade, aprendizagem, área de educação e formação, docente, educação formal, ensino, retenção, transição, conclusão, oferta de

educação e formação, oferta formativa, oferta educativa, sistema de educação, sistema educativo, sistema de educação e formação.

Os principais constrangimentos prenderam-se com: (1) existência de perspetivas frequentemente não convergentes, no âmbito da Educação e da Formação; (2) proliferação, inconstância, disparidade e incoerência terminológica e semântica da documentação legislativa, normativa e técnica; (3) diferenças legislativas existentes entre o Continente e as Regiões Autónomas.

Todo o trabalho efetuado possibilitou a elaboração de uma lista final com um conjunto de 327 conceitos para entrar em vigência no SEN na área de Educação, Formação e Aprendizagem. Esta nova lista de conceitos foi aprovada pela 53.^a deliberação da Secção Permanente de Coordenação Estatística, publicada com o n.º 327/2017, Diário da República n.º 82/2017, Série II de 2017-04-27.

3.2 Análise e impacto dos Processos RVCC

A definição do conceito de aluno, e mais concretamente a sua ligação aos Processos de Reconhecimento, Validação e Certificação de Competências (Processos RVCC), foi dos temas mais debatidos.

Neste sentido, e antes de se entrar na problemática dos RVCC, é importante esclarecer o conceito de aluno. Este conceito, na sua forma final, considera que aluno é um “Indivíduo que, após um ato de registo administrativo, participa em percursos de educação e formação no âmbito da educação formal”, sendo que por educação formal se entende a “educação intencional, institucionalizada e planeada que se materializa em oferta de educação e formação, confere certificação escolar ou dupla certificação, apresenta uma sucessão progressiva de níveis de escolaridade e é ministrada por entidades públicas ou privadas reconhecidas pelas autoridades nacionais competentes em matérias de educação e formação”. Nesse sentido, excluem-se da educação formal: a formação profissional e técnica nas empresas; a formação exclusivamente em contexto de trabalho; a formação sem reconhecimento formal das autoridades nacionais competentes e os programas de curta duração de menos de um semestre ou duração

equivalente a tempo completo, segundo legislação em vigor.

De acordo com o definido no manual da ISCED 2011, a educação é um processo de comunicação organizado e concebido para suscitar aprendizagens. Também o Manual do UOE [5] remete o conceito de educação para o definido pela ISCED, destacando que as palavras-chave para o seu entendimento são: comunicação (transferência de conhecimentos), aprendizagem (aquisição de), organização (das aprendizagens através de programas e dos próprios docentes), sustentabilidade (a experiência de aprendizagem tem elementos de duração e continuidade).

Subentende-se, assim, que um “aluno” faz uma aprendizagem sustentada de conhecimentos formativos organizados de forma científica e pedagógica, durante um período estabelecido e reconhecido como necessário do ponto de vista educativo, findo o qual lhe é conferido um determinado grau da classificação educativa, caso tenha provado que assimilou estes conhecimentos.

Ora, em relação aos “alunos” em processos RVCC, a problemática assentava na resposta a dois problemas principais: a forma de inclusão dos Processos RVCC no esquema conceptual associado ao referido conjunto de conceitos; a equiparação - ou não - dos adultos em Processos RVCC a alunos, nos processos de produção, reporte e difusão de informação estatística.

Em Portugal, os Processos RVCC partem da valorização dos conhecimentos previamente adquiridos pelos alunos adultos em contextos formais, não formais ou informais. Por outro lado, tal como em outras ofertas de educação e formação, a carga de esforço dos alunos adultos, para a conclusão com êxito do Processo RVCC, não se reduz à simples presença e participação nas ações de formação promovidas no âmbito do Processo, mas igualmente ao trabalho individual desenvolvido. Para que os adultos em RVCC possam ser correntemente classificados como alunos, a duração teórica dos processos deverá ser igual ou superior a um semestre letivo. A tabela seguinte serve para caracterizar alguns dos principais indicadores administrativos relacionados com esta temática.

Como se pode observar na tabela 1, as maiores parcelas de cer-

Tabela 1: Processos RVCC em Portugal, entre os anos letivos 2009/2010 e 2014/2015

	Adultos certificados (1)	Dias ações (2)	Horas processo (3)	Formação complementar (4)
Total	264.273	283	55	21
Básico	169.376	225	56	23
1.º Ciclo	986	168	52	23
2.º Ciclo	19.151	190	58	25
3.º Ciclo	149.239	230	56	22
Secundário	94.897	384	52	18

Fonte: SIGO, Sistema de Informação e Gestão da Oferta Educativa

Notas: (1) Número de alunos/adultos que concluíram, com êxito, o Processo RVCC; (2) Número médio de dias em que se desenvolveu o Processo RVCC; (3) Número médio de horas de desenvolvimento de todas as ações do Processo RVCC; (4) Número médio de horas de formação complementar.

tificação de adultos via conclusão com êxito de Processos RVCC registam-se no 3.º ciclo do ensino básico (56,5%) e no ensino secundário (35,9%). Por outro lado, a duração efetiva média de desenvolvimento de um Processo RVCC foi de 283 dias (225 dias, no ensino básico; 384 dias, no ensino secundário; comportamento crescente, com o nível de ensino e o ciclo de estudos). Em termos dinâmicos, é importante notar que o número de alunos/adultos matriculados em RVCC tem diminuído de forma acentuada, sendo atualmente muito reduzido.

As perspectivas que se colocavam eram, então, as seguintes:

1) Consideração dos Processos RVCC como uma Oferta de Educação e Formação, orientada para adultos.

Esta opção conduziria à contabilização dos adultos em Processos RVCC como alunos, quer no que se refere a “Matrículas”, quer no que se refere a “Resultados escolares”.

2) Consideração dos adultos em Processos RVCC como estando numa modalidade de Qualificação.

O que em termos práticos se traduz na consideração dos adultos em Processos RVCC apenas no âmbito dos “Resultados escolares”.

Importa igualmente verificar o impacto da decisão de equiparar, ou não, os adultos em Processos RVCC a alunos. A não equiparação, teria um efeito negativo de cerca de 200 mil alunos (em 2009), 190 mil (em 2010) e 10.500 (em 2016) (Figura 1).

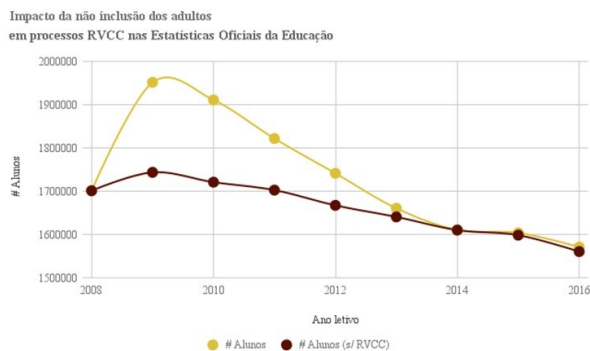


Figura 1: Impacto da não inclusão dos adultos em processos RVCC nas Estatísticas Oficiais da Educação

Perante a análise efetuada, foi decidido, em termos gerais, aprovar o conceito de aluno nos moldes referidos anteriormente e integrar os Processos RVCC na área das “Ofertas da Educação e Formação”, na Estrutura do Sistema Conceptual da proposta de conceitos [6].

4 Classificação Internacional Tipo da Educação (CITE)

A *International Standard Classification of Education (ISCED)* 2011 [7] foi formalmente aprovada em 2011 pela UNESCO, com o objetivo

de substituir a versão anterior (ISCED 97). A Classificação Internacional Tipo Educação (CITE) 2011 constitui a versão portuguesa da ISCED e a sua implementação foi aprovada pela 52.^a Deliberação da Secção Permanente de Coordenação Estatística, publicada com o n.º 343/2017, Diário da República n.º 84/2017, Série II de 2017-05-02. A CITE 2011, que passou a ser de carácter obrigatório no reporte internacional dos dados a partir de 2014, constitui o quadro de referência que permite a apresentação normalizada de estatísticas muito diversas e relevantes para a elaboração das políticas relativas à educação, de acordo com um conjunto de definições e de conceitos acordados internacionalmente, garantindo assim a comparabilidade, ao nível internacional, dos indicadores obtidos.

A sua correta aplicação possibilitará a existência de uma base comparativa das diferentes ofertas existentes nos Sistemas de Educação e Formação de cada país membro da UNESCO, da OCDE e do Eurostat. No âmbito do mandato do GTEEF, referido no ponto 1, o desenvolvimento dos trabalhos a nível nacional, incidiu sobre duas vertentes: tradução da ISCED 2011 para a língua portuguesa; constituição do quadro de equivalências entre os níveis ISCED 97, ISCED 2011 e os níveis de escolaridade vigentes no sistema de educação e formação português. Como resultado final do trabalho desenvolvido, sintetiza-se abaixo a relação entre os diferentes níveis da CITE 2011 e os níveis e programas de ensino em Portugal:

- CITE 0: Educação pré-escolar;
- CITE 1: Ensino básico - 1.º ciclo e 2.º ciclo;
- CITE 2: Ensino básico - 3.º ciclo
- CITE 3: Ensino secundário
- CITE 4: Ensino pós-secundário, não superior
- CITE 5: Ensino superior - curso técnico superior profissional
- CITE 6: Ensino superior - bacharelato e licenciatura de 1.º ciclo de Bolonha

- CITE 7: Ensino superior - licenciatura pré-Bolonha, mestrado integrado de Bolonha, mestrado de 2.º ciclo de Bolonha, mestrado pré-Bolonha
- CITE 8: Ensino superior - doutoramento de 3.º ciclo de Bolonha e doutoramento pré-Bolonha.

4.1 Principais eixos orientadores da CITE

Por comparação com a anterior classificação, ao nível dos principais eixos orientadores da CITE 2011, há três aspetos essenciais a destacar:

- a) apresenta mais dois níveis a um dígito, passando de 7 para 9 níveis (o ensino superior passa a ter 4 níveis face a 2 níveis na versão anterior);
- b) apresenta um maior detalhe, podendo ir a um nível de desagregação até 3 dígitos (o segundo dígito respeita à orientação do programa - geral ou vocacional, no ensino não superior e no ensino superior de curta duração; académico e profissional, no ensino superior - e o terceiro dígito refere-se à etapa e duração dentro do nível);
- c) permite uma codificação em paralelo dos programas educativos e do nível de escolaridade completo.

No que particularmente se refere à alínea c), importa compreender a diferença entre: 1) por um lado, a classificação dos programas educativos dos vários níveis de ensino - essencialmente, a classificação de dados administrativos e/ou dados estatísticos recolhidos pela DGEEC relativos a matrículas/inscrições, resultados escolares e recursos humanos (pessoal docente e pessoal não docente); 2) por outro, a classificação do nível de escolaridade completo- avaliação dos stocks populacionais em termos de nível de escolaridade concluída com sucesso, a partir de dados estatísticos recolhidos pelo INE e por

outras autoridades estatísticas, através dos Censos e de diferentes inquéritos às famílias e às empresas.

Durante o processo de implementação da CITE 2011, surgiram problemas de classificação de programas educativos ou níveis de escolaridade considerados de fronteira, por exemplo, situados entre os níveis 4 e 5 da CITE, entre os níveis 5 e 6 e entre os níveis 6 e 7. Através de um longo debate entre as diferentes organizações internacionais, nomeadamente, a UNESCO, o Eurostat e a OCDE, e os Estados Membros, foi então sublinhada a necessidade de considerar três critérios fundamentais para a distinção entre os programas educativos e níveis de escolaridade:

- requisitos de acesso - a conclusão bem-sucedida dos programas permite ou não o acesso a programas classificados em níveis da CITE mais elevados;
- conteúdos - os sucessivos níveis CITE refletem graus crescentes de complexidade e de especialização dos programas educativos;
- duração - proxy ao critério da complexidade dos programas educativos referido no ponto 2.

Como se irá verificar nos pontos seguintes, foi com estas premissas que as decisões relativas às classificações dos programas em Portugal foram tomadas.

4.2 Programas de estudo na fronteira de dois níveis da CITE

De entre as análises técnicas desenvolvidas destacam-se, pela relevância em termos do número de alunos matriculados e do número de indivíduos detentores de um nível de escolaridade, as seguintes:

- Análise transversal da classificação dos programas de ensino superior pré e pós-Bolonha, nomeadamente nos aspetos que resultam da redução generalizada da duração teórica dos cursos;

- Classificação dos Cursos de Especialização Tecnológica (CET) e Cursos Técnicos Superiores Profissionais (TeSP), no âmbito da Classificação CITE 2011;
- Classificação das Licenciaturas pré e pós-Bolonha, igualmente no âmbito da classificação CITE 2011.

4.2.1 Programas pré e pós-Bolonha

A problemática prendia-se com a redução da duração teórica dos programas que conferem os graus de licenciado, mestre e doutor resultante da implementação do processo de Bolonha. Em causa estavam então os cursos de bacharelato, licenciatura e mestrado (pré-Bolonha e pós-Bolonha). Existindo duas perspetivas diferentes, a solução passaria pela discussão em torno ou da adoção de um quadro de equivalência com transposição direta da CITE 97 para a CITE 2011, privilegiando o critério diploma; ou privilegiando a duração teórica como forma de medir a complexidade de um curso, e, portanto, o nível da Classificação CITE em que esse curso seria colocado. Como solução decidiu-se pela harmonização internacional transversal (tanto quanto possível) na classificação de programas com características similares nos diferentes países.

4.2.2 CET e TeSP - CITE 4 ou CITE 5?

Após uma ampla discussão aos níveis nacional e internacional (Eurostat/UNESCO/OCDE/Estados Membros) foi estabelecido um acordo de princípio em que se considerava que programas com duração de 1 a 3 semestres deveriam ser classificados no nível 4 da CITE e programas com duração de 4 a 5 semestres deveriam ser classificados no nível 5 da CITE. Neste sentido, os programas e qualificações associados aos Cursos de Especialização Tecnológica (CET) foram classificados no nível 4 da CITE 2011 e os programas e qualificações associados aos Cursos Técnicos Superiores Profissionais (TeSP) foram classificados no nível 5 da CITE 2011.

4.2.3 Licenciaturas pré e pós-Bolonha

De todas as discussões técnicas que existiram, as que envolveram a classificação de Licenciaturas pré-Bolonha e de Licenciaturas pós-Bolonha (Licenciaturas de 1.º ciclo) foram as mais complexas. Primeiro importa descrever as características específicas de cada um dos programas:

- Em termos da duração teórica - 3 a 4 anos pós-Bolonha e 4 a 6 anos pré-Bolonha (diminuição na duração de 1 a 2 anos);
- Em termos do diploma/grau obtido - qualquer destas ofertas educativas atribui o mesmo diploma (licenciatura) e o mesmo grau (licenciado);
- Em termos da classificação na CITE 97: 5A, independentemente de se tratar de licenciaturas pré-Bolonha ou pós-Bolonha.

Depois, a grande questão: Classificar ambas as Licenciaturas - pré e pós Bolonha - no mesmo nível CITE 6, “*Bachelor or equivalent*”? Ou classificar de forma diferenciada, a licenciatura pós-Bolonha como “*Bachelor or equivalent*” (CITE 6) e licenciatura pré-Bolonha como “*Master or equivalent*” (CITE 7)?

É importante notar que qualquer opção que se viesse a seguir teria diferentes implicações. Por exemplo, a classificação de ambas as licenciaturas no mesmo nível CITE, privilegiando como critério de decisão o grau atribuído, permitiria fazer o seguimento ao longo do tempo da percentagem de licenciados na população ativa, tal como “historicamente” se vinha a fazer. Em contrapartida, em termos comparativos internacionais, o país surgiria como tendo uma repartição pelos níveis 6 (*Bachelor or equivalent*) e 7 (*Master or equivalent*) com um maior peso no nível mais baixo, o que não traduziria o “real valor”, o nível de conhecimento e preparação da população ativa portuguesa. Em particular, todos os licenciados pré-Bolonha com licenciaturas, que passaram posteriormente a mestrados integrados (como é o caso, por exemplo, das engenharias, medicina, ciências

farmacêuticas, arquitetura, ...) seriam classificados como “*Bachelor or equivalent*”.

Na secção seguinte apresentam-se os resultados da análise de impacto de cada uma das opções. Como resultado e tendo, ainda, como base outros pareceres técnicos, mais concretamente orientações internacionais, decidiu-se pela classificação diferenciada de acordo com o critério de duração: as licenciaturas pós-Bolonha ficaram classificadas no nível 6 da CITE 2011 e licenciaturas pré-Bolonha no nível 7 da CITE 2011.

5 Licenciaturas pré e pós-Bolonha - impacto de cada opção de classificação

Para analisar o real impacto das decisões técnicas que iriam ser assumidas, começou-se por estimar o número de diplomados com formações de nível superior de 5 ou mais anos vs. os de 4 ou menos anos, na população ativa. Para este exercício, recorreu-se a duas bases de dados: Dados dos censos 2011, organizados por idade dos indivíduos, nível de formação superior (Bacharelato, Licenciatura, Mestrado e Doutoramento) e área de educação e formação (CNAEF); Dados do RAIDES (diplomados por nível de formação superior e por idade entre 2011 e 2014).

As principais decisões metodológicas, apresentadas na figura anterior, assentaram nos seguintes pressupostos:

- todos os diplomados com mestrado ou doutoramento integraram o grupo “formações com 5 ou mais anos”;
- todos os diplomados com Bacharelato integraram o grupo “formações com 4 ou menos anos”;
- todos os diplomados com licenciatura em escalões etários a partir dos 52 anos (em 2011) integraram o grupo “formações com 5 ou mais anos”;

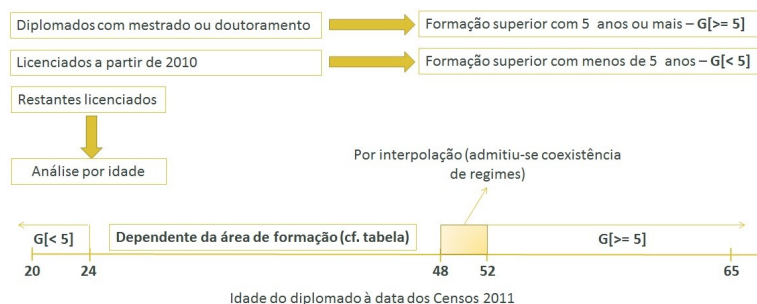


Figura 2: Metodologia de estimação do número de diplomados com formação de 5 ou mais anos ($G[\geq 5]$) e de 4 ou menos anos ($G[\leq 4]$)

- os diplomados com idades entre os 48 e 51 anos foram repartidos, gradualmente, pelos grupos “formações com 5 ou mais anos” e “formações com 4 ou menos anos” (uma vez que são idades de charneira para os licenciados no período 1988 a 2006 e anterior a 1988);
- dos diplomados com licenciatura cujas idades variam entre 24 e 47 anos, foram integrados no grupo “formações com 5 ou mais anos” a parte correspondente a áreas onde as licenciaturas vigentes no período 1988 a 2006 revelavam longas durações.

Admitindo como consensual que os licenciados cujas licenciaturas têm 5 ou mais anos de duração deverão ser contabilizados no nível CITE 7, a dificuldade colocava-se principalmente no período de 1988 a 2006 onde coexistiram licenciaturas de 4, 5 e 6 anos. Nesse período, as licenciaturas de 5 ou mais anos eram, tipicamente, as das grandes áreas com ordens profissionais (engenharias, medicina, farmácia, arquitetura, direito, psicologia) e ainda as de ensino para os 2.º e 3.º ciclos do ensino básico e para o ensino secundário.

Em outras áreas também existiam, à data, licenciaturas de 5 anos (nomeadamente, em informática, ciências sociais, física, matemá-

tica, geologia) mas, por não haver informação sistematizada quanto à respetiva duração, optou-se por repartir equitativamente os diplomados nessas áreas por cada um dos grupos ($G[\geq 5]$ e $G[< 5]$). Para todas as restantes áreas os diplomados contabilizaram para o grupo $G[< 5]$.

Chegou-se então às seguintes estimativas: 11% da população com idades (em 2014) compreendidas entre os 25 e os 64 anos teria um curso superior de duração inferior a 5 anos e 12% teria um curso superior de duração igual ou superior a 5 anos.

Os gráficos da Figura 3 ilustram a distribuição dos níveis de classificação CITE 6, 7 e 8 na população portuguesa consoante se atribua o nível 6 a todas as licenciaturas ou se atribua o nível 6 apenas às licenciaturas pós-Bolonha e o nível 7 às licenciaturas pré-Bolonha. De acordo com o Cenário 1, Portugal surgiria nos reportes nacionais e internacionais como tendo apenas 4,5% da população dos 25 aos 64 anos com o equivalente ao grau de mestre ou superior (quando as estimativas acima apresentadas indicam que essa percentagem seja, efetivamente, de 12%) enquanto que no Cenário 2, a referida percentagem fica mais próxima (17,1%), embora, naturalmente, sobreavaliada.

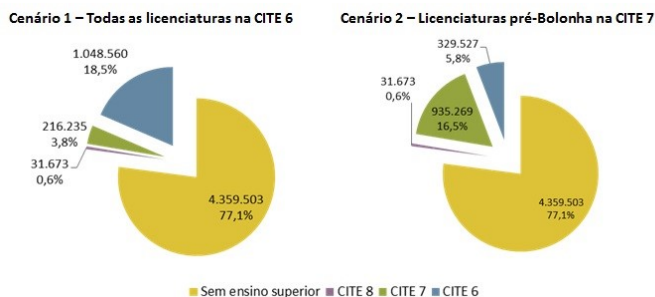


Figura 3: Número de diplomados nas CITE 6, 7 e 8 na população com 25 a 64 anos (em 2015), de acordo com os Cenários 1 e 2

Na segunda parte do exercício, admitindo a estacionariedade nas taxas de conclusão das licenciaturas e mestrados e nas taxas de prosseguimento para mestrado entre os licenciados, de acordo com os indicadores mais recentes, previu-se sucessivamente ao longo dos anos a proporção da população com cada tipo de diploma no grupo etário dos 25 aos 64 anos.

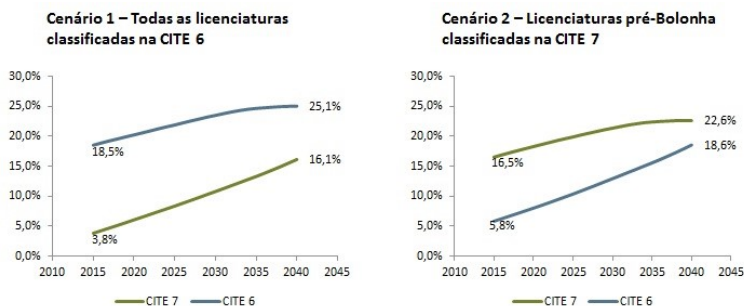


Figura 4: Previsão da evolução da percentagem de diplomados na população [25 a 64 anos] em cada nível de CITE, de acordo com os Cenários 1 e 2

Nas premissas do Cenário 1, de classificação de todas as licenciaturas no nível 6 da CITE, parte-se em 2015 de uma repartição [CITE 7:CITE 6] de [3,8%:18,5%], havendo uma previsão de crescimento em ambos os níveis da ISCED até 2030, com pequeno decréscimo na CITE 6 e um crescimento mais acelerado na ISCED 7 a partir de 2030.

Nas condições do Cenário 2, de classificação das Licenciaturas pré-Bolonha no nível 7 da CITE, parte-se em 2015 de uma repartição [CITE 7:CITE 6] de [16,5%:5,8%], havendo também aqui uma previsão de crescimento em ambos os níveis CITE até 2030, neste caso com um pequeno decréscimo na ISCED 7 (fenómeno que poderá eventualmente não se verificar caso venha a haver um aumento da

taxa de prosseguimento para mestrado entre os licenciados) e um crescimento continuado na ISCED 6 a partir de 2030.

Referências

- [1] Eurostat (2011), *Código de Conduta para as Estatísticas Europeias*, tradução realizada pelo INE, IP., Lisboa.
- [2] *Lei do Sistema Estatístico Nacional*, Lei 22/2008 (D.R. 921^a Série, de 2008-05-13).
- [3] Conselho Superior de Estatística (2017), *Relatório do Grupo de Trabalho de Estatísticas da Educação e Formação*, Lisboa.
- [4] Conselho Superior de Estatística (2017), *Relatório do subgrupo “Conceitos de Produção e Difusão Estatística nas Áreas da Educação e Formação”, do Grupo de Trabalho de Estatísticas da Educação e Formação*, Lisboa
- [5] UNESCO-UIS/OECD/EUROSTAT (2010), *Data Collection on Education Statistics Manual*, Montreal, Paris, Luxemburgo.
- [6] 53.^a *Deliberação da Secção Permanente de Coordenação Estatística*, publicada em Diário da República com o n.º 327/2017, Diário da República n.º 82/2017, Série II de 2017-04-27
- [7] UNESCO [2012], *International Standard Classification of Education*, Montreal, Canadá

Omissões e dimensão da amostra: Impacto sobre medidas de qualidade do ajustamento em modelos de análise fatorial confirmatória

Paula C.R. Vicente

ULHT-Escola de Ciências Económicas e das Organizações e Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, *paula.vicente@ulusofona.pt*

Maria de Fátima Salgueiro

Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, *fatima.salgueiro@iscte-iul.pt*

Palavras-chave: Dados Omissos; Estudo de Monte Carlo; Modelo de Análise Fatorial Confirmatória; Mplus.

Resumo: Neste estudo de simulação é analisado o impacto nas medidas de qualidade de ajustamento rácio χ^2/gl , SRMR e RMSEA, da dimensão da amostra e da existência de não respostas, em modelos de análise fatorial confirmatória. São consideradas omissões planeadas pelo investigador e não planeadas. Os índices que se mostraram mais afetados pela existência de não respostas foram o rácio χ^2/gl e o SRMR. Em amostras de maior dimensão os índices analisados apresentaram melhores valores.

1 Introdução

A utilização de modelos com equações estruturais, em particular, modelos de análise fatorial confirmatória (AFC) (Bollen [1], Salgueiro [5]), tem revelado uma importância crescente em diversas áreas das ciências sociais. Todavia, a impossibilidade de dispor de amostras de dimensão razoável face ao tipo de modelação desejada

e a existência de não respostas nas amostras recolhidas, são circunstâncias comuns que se traduzem em limitações e/ou desafios em termos estatísticos e metodológicos. Por outro lado, apesar de uma das grandes vantagens do *framework* dos modelos com equações estruturais ser a possibilidade de aferir da qualidade do ajustamento modelo-dados, face à variedade de medidas disponíveis na literatura e implementadas nos pacotes estatísticos existentes, é recomendada a utilização de medidas de diferentes tipos/famílias, uma vez que os valores para elas obtidos podem não ser concordantes em termos da qualidade do ajustamento do modelo em análise.

Na área das ciências sociais muitos estudos utilizam dados recolhidos através de inquérito, sendo comum a existência de não respostas nos dados obtidos desta forma. As omissões podem ser *item non response*, caso em que o indivíduo não responde a uma ou mais perguntas, ou *unit non response*, caso em que o indivíduo não responde a todas as perguntas. Por outro lado, as omissões podem também ocorrer de forma intencional e de acordo com a vontade do investigador, situação que se designa por *planned missing design* (PMD) (Enders, [2]). Este procedimento tem por objetivo aumentar a qualidade dos dados evitando o esforço de inquirição e o consequente abandono do questionário por parte dos respondentes, que é uma das principais razões para a existência de não respostas em dados recolhidos por questionário.

Este trabalho apresenta os principais resultados de um estudo de simulação realizado recorrendo ao pacote estatístico Mplus 7 (Muthén e Muthén [4]), no qual se pretende aferir o efeito da existência de dados omissos, bem como o efeito dimensão da amostra, nas medidas de qualidade do ajustamento modelo-dados. É considerado um modelo de AFC com dois fatores, medidos por dois, três ou quatro indicadores, são considerados diferentes valores para os pesos fatoriais e diferentes graus de correlação entre os dois fatores. São consideradas e analisadas as seguintes medidas: rácio da estatística do qui-quadrado pelos graus de liberdade (χ^2/gl), *Standardized Root Mean Square Residual* (SRMR) e *Root Mean Square Error of Approximation* (RMSEA). São usadas amostras de diferentes dimensões,

com diferentes padrões e percentagens de omissão.

2 Metodologia

2.1 Modelo de Análise Fatorial Confirmatória

O modelo de Análise Fatorial Confirmatória corresponde ao modelo de medida de um modelo de equações estruturais e representa a relação entre as variáveis observadas e as variáveis latentes do modelo. A equação do modelo de AFC é:

$$\mathbf{X} = \mathbf{\Lambda}_X \boldsymbol{\xi} + \boldsymbol{\delta} \quad (1)$$

em que, $\mathbf{X}^T = (X_1, X_2, \dots, X_k)$ é o vetor das k -variáveis observadas, $\boldsymbol{\xi}^T = (\xi_1, \xi_2, \dots, \xi_q)$ é o vetor das q variáveis latentes, $\boldsymbol{\delta}^T = (\delta_1, \delta_2, \dots, \delta_k)$ é o vetor dos k -termos residuais do modelo de medida e $\mathbf{\Lambda}_X$ é a matriz dos pesos fatoriais. Os termos residuais não estão correlacionados com as variáveis latentes e seguem uma distribuição normal multivariadas com média zero e matriz de variâncias-covariâncias $\boldsymbol{\Theta}_\delta$. Tal implica que, dados gerados a partir de um modelo definido pela equação 1 sigam distribuição normal multivariada. A matriz de variâncias-covariâncias implícita ao modelo, $\boldsymbol{\Sigma}$, é dada por:

$$\boldsymbol{\Sigma} = \mathbf{\Lambda}_X \boldsymbol{\Phi} \mathbf{\Lambda}_X^T + \boldsymbol{\Theta}_\delta, \quad (2)$$

sendo, $\boldsymbol{\Phi}$ a matriz de variâncias-covariâncias das variáveis latentes $\boldsymbol{\xi}$ e $\boldsymbol{\Theta}_\delta$ a matriz diagonal de variâncias-covariâncias dos termos residuais.

Estimar um modelo de AFC consiste em encontrar valores para os parâmetros do modelo que resultem numa matriz de variâncias-covariâncias que reproduza o melhor possível a matriz de variâncias-covariâncias implícita ao modelo teórico considerado. Assim, sendo $\boldsymbol{\Sigma}$ a matriz de variâncias-covariâncias subjacente ao modelo de AFC, pretende-se que a matriz de variâncias-covariâncias entre as variáveis observadas \mathbf{S} , esteja o mais próximo possível da matriz $\boldsymbol{\Sigma}$. Os

valores estimados dos parâmetros do modelo (λ_{ij} e ϕ_{ij}) são determinados por forma a minimizar uma função distância, F , entre $\hat{\Sigma}$ e \mathbf{S} .

A função, F , é definida de acordo com o método de estimação considerado, sendo definida por

$$F = \text{tr}[\mathbf{S}\Sigma^{-1}] + \log|\Sigma| - \log|\mathbf{S}| - k, \quad (3)$$

quando é considerado o método da máxima verosimilhança.

2.2 Medidas de Qualidade do Ajustamento

Entende-se por qualidade do ajustamento modelo-dados o grau em que o modelo especificado e estimado reproduz a estrutura de variâncias-covariâncias (ou de correlações) observada na amostra. Na literatura sobre modelos com equações estruturais é possível encontrar um conjunto alargado de medidas (de diferentes tipos), que são usualmente calculadas com o objetivo de avaliar o ajustamento modelo-dados. Todavia, de acordo com Schumacker e Lomax [6], existem três índices de ajustamento que devem ser sempre apresentados, qualquer que seja o modelo considerado: o rácio da estatística de qui-quadrado pelos respetivos graus de liberdade (χ^2/gl), o SRMR e o RMSEA.

2.2.1 Estatística de χ^2 e o rácio χ^2/gl

A estatística de qui-quadrado é um teste estatístico cujo objetivo é a não rejeição da hipótese nula de que a estrutura de variâncias-covariâncias entre as variáveis observadas reproduz a estrutura de variâncias-covariâncias do modelo especificado. Esta estatística é definida como:

$$\chi^2 = -2 \left\{ -1/2(n-1) \left[\text{tr}(\mathbf{S}\hat{\Sigma}^{-1}) + \log|\hat{\Sigma}| - \log|\mathbf{S}| - k \right] \right\}, \quad (4)$$

em que n é a dimensão da amostra e k o número de variáveis. Este teste assume a normalidade multivariada dos dados e é sensível à dimensão da amostra. Assim, quando a dimensão da amostra aumenta

a estatística χ^2 tende a aumentar (Bollen [1]). O rácio χ^2/gl é um índice obtido dividindo a estatística de qui-quadrado pelos respetivos graus de liberdade, sendo que um valor inferior a 2 ou 3 indica um bom ajustamento, embora estes valores não sejam consensuais na literatura (Salgueiro [5]).

2.2.2 Standardized Root Mean Square Residual (SRMR)

O índice SRMR é uma medida baseada na média dos resíduos estandardizados entre a matriz de variâncias-covariâncias observada e a que está subjacente ao modelo. Este índice calcula-se através de:

$$SRMR = \sqrt{\frac{\sum_{i=1}^p \sum_{j=1}^l [(s_{ij} - \hat{\sigma}_{ij})/(s_{ii}s_{jj})]^2}{k(k+1)/2}}, \quad (5)$$

em que, s_{ij} e $\hat{\sigma}_{ij}$ são elementos da matriz \mathbf{S} e $\hat{\Sigma}$, respetivamente. Este índice indica um bom ajustamento modelo-dados quando o seu valor é inferior a 0.05 (Salgueiro [5]).

2.2.3 Root Mean Square Error of Approximation(RMSEA)

O índice RMSEA é uma medida baseada na diferença entre a matriz de variâncias-covariâncias pelos graus de liberdade e a matriz que está subjacente ao modelo. Este índice calcula-se através de:

$$RMSEA = \sqrt{\max \left\{ \left(\frac{F(S, \hat{\Sigma})}{\nu} - \frac{1}{n-1} \right), 0 \right\}}, \quad (6)$$

em que, $F(S, \hat{\Sigma})$ é a função distância, definida em função do método de estimação, e ν são os graus de liberdade.

Este índice produz melhores valores quando a dimensão da amostra é grande, pois o termo $1/(n-1)$ tende para zero quando n tende para infinito (ver equação 6). O RMSEA também tem em consideração o número de indicadores, dado que este número influencia os graus de liberdade do modelo.

De acordo com Salgueiro [5], valores de RMSEA inferiores a 0.05 indicam um bom ajustamento modelo-dados, mas valores até 0.08 podem ser considerados aceitáveis.

2.3 Opções do Estudo de Simulação

Pode realizar-se um estudo de simulação utilizando procedimentos de Monte Carlo recorrendo ao pacote estatístico Mplus (Muthén e Asparouhov [3]). Este *software* permite gerar m amostras de dados a partir da estrutura de um determinado modelo (neste estudo, um modelo AFC), cujos parâmetros populacionais são definidos à priori pelo investigador. Para cada uma das m amostras geradas, e de uma forma, integrada, é estimado um modelo AFC, obtendo-se deste modo, m estimativas para cada uma das medidas de qualidade de ajustamento. Para cada uma das medidas disponibilizadas pelo software, é apresentado um valor médio calculado a partir das m amostras independentes que foram geradas, assim como, o respetivo desvio padrão. Para além da média das várias medidas de ajustamento modelo-dados disponibilizadas pelo Mplus, também são fornecidas as médias e os erros padrão das estimativas dos parâmetros, assim como o erro quadrático médio, a cobertura e a potência do teste (Muthén e Muthén [4], Vicente e Salgueiro [7]). O método de estimação utilizado é o da máxima verosimilhança, mas quando as amostras geradas apresentam omissões é considerado o método da máxima verosimilhança de informação completa.

Neste estudo de simulação foram geradas, a partir da estrutura de um modelo de AFC com dois fatores medidos por dois, três e quatro indicadores cada, 1000 amostras de dados, com 100, 250 e 500 observações cada (ver exemplo na figura 1). Foram utilizados como valores dos parâmetros populacionais, $\lambda = 0,8$ e $\lambda = 0,6$, o primeiro corresponde a um peso fatorial considerado bom e o outro um valor no limite do aceitável, sendo os correspondentes valores de fiabilidade dos indicadores de, respetivamente, 0,64 e 0,36. Para a correlação entre fatores latentes foram considerados os valores $\phi_{21} = 0,1$, 0,25 e 0,5.

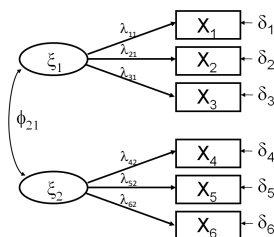


Figura 1: Modelo de análise fatorial confirmatória com dois fatores, ξ_1 e ξ_2 , correlacionados entre si, cada um deles medido por três indicadores

Além de serem geradas amostras com dados completos, são ainda geradas amostras com omissões. Amostras com 25% de omissões em cada indicador, de acordo com um mecanismo completamente aleatório (que designamos por omissões MCAR) e amostras com omissões planeadas pelo investigador (que designamos por omissões PMD), tal como ilustrado na tabela 1. A tabela 1, apresenta o caso, em que existem 25% de observações com não respostas nos indicadores do factor 1, X_1 , X_2 e X_3 , as restantes 75% das observações não apresentam omissões em qualquer dos indicadores do modelo. De referir que, quando se considera omissões nos dados, apenas são usados modelos de AFC com três ou quatro indicadores em cada fator, com pesos fatoriais no limite do aceitável (0,6). Modelos em que cada fator é medido por dois indicadores não foram considerados, uma vez que constituem o caso mais extremo.

3 Resultados

Para cada um dos índices em análise foi calculado um intervalo de confiança a 95% e é apresentado em cada uma das diferentes situações, o seu limite superior.

indicadores	X_1	X_2	X_3	X_4	X_5	X_6
75% das observações	O	O	O	O	O	O
25% das observações	×	×	×	O	O	O

Tabela 1: Desenho das omissões PMD para o modelo AFC com 3 indicadores em cada fator. O -valor observado, × - valor omisso

3.1 Rácio χ^2/gl

A análise da tabela 2 permite verificar que o limite superior do intervalo a 95% para o rácio χ^2/gl diminui quando a dimensão da amostra aumenta e mantém-se quando o grau de correlação entre fatores latentes aumenta, em modelos com três ou quatro indicadores por fator. Todavia, em modelos em que cada fator é medido por dois indicadores, este valor aumenta quando aumenta a dimensão da amostra e quando aumenta o valor da correlação entre fatores latentes. De salientar que, em modelos deste tipo e se a correlação entre fatores latente é forte (0,5), então o rácio χ^2/gl pode assumir valores próximos de quatro, o que indica um mau ajustamento modelo-dados, qualquer que seja o valor assumido pelos pesos fatoriais.

Por outro lado, é possível verificar que o limite superior do intervalo a 95% para este rácio não revela ser afetado pelo valor assumido pelos pesos fatoriais, em modelos com três ou quatro indicadores. Aliás, a análise da tabela 2 permite verificar que o valor do rácio tende a estabilizar.

A tabela 3 mostra que, em modelos em que cada fator é medido por três indicadores ou quatro indicadores, os valores aumentam se existem 25% de omissões em cada indicador (MCAR) e se as omissões são PMD os valores são próximos, quando comparados com os resultados dos modelos sem omissões.

		2 indicadores		3 indicadores		4 indicadores	
		$\lambda = 0,6$	$\lambda = 0,8$	$\lambda = 0,6$	$\lambda = 0,8$	$\lambda = 0,6$	$\lambda = 0,8$
$\phi_{12} = 0,1$	n=100	1,787	1,993	2,105	2,153	1,702	1,709
	n=250	1,995	2,163	2,073	2,074	1,677	1,675
	n=500	2,180	2,581	1,998	2,001	1,627	1,625
$\phi_{12} = 0,25$	n=100	2,285	2,728	2,115	2,151	1,702	1,708
	n=250	2,836	3,446	2,065	2,071	1,675	1,674
	n=500	3,326	3,724	1,992	2,000	1,626	1,624
$\phi_{12} = 0,5$	n=100	3,475	3,903	2,116	2,144	1,701	1,707
	n=250	3,919	4,029	2,046	2,060	1,666	1,670
	n=500	3,975	3,969	1,971	1,991	1,626	1,621

Tabela 2: Limite superior de um intervalo de confiança a 95% para o rácio χ^2/gl , modelo AFC medido por 2, 3 ou 4 indicadores e dados completos

		3 indicadores		4 indicadores	
		MCAR	PMD	MCAR	PMD
$\phi_{12} = 0,1$	n=100	2,157	2,074	1,809	1,766
	n=250	2,085	2,062	1,710	1,713
	n=500	2,037	1,997	1,631	1,643
$\phi_{12} = 0,25$	n=100	2,149	2,075	1,810	1,767
	n=250	2,079	2,049	1,709	1,711
	n=500	2,031	1,990	1,629	1,643
$\phi_{12} = 0,5$	n=100	2,153	2,096	1,811	1,769
	n=250	2,065	2,029	1,709	1,703
	n=500	2,025	1,975	1,631	1,644

Tabela 3: Limite superior de um intervalo de confiança a 95% para o rácio χ^2/gl , modelo AFC medido por 3 ou 4 indicadores, $\lambda = 0,6$ e dados com omissões MCAR e PMD

3.2 SRMR

A análise da tabela 4 permite verificar que, os valores obtidos do índice SRMR não são afetados pelo aumento da correlação entre os fatores latentes do modelo de AFC. Todavia, apresenta piores resultados quando os pesos fatoriais no modelo considerado são menores, exceto no caso de o modelo ter dois fatores medido por dois indicadores cada.

Os valores obtidos mostram, ainda que, aumentar a dimensão da amostra provoca a diminuição deste índice, mas quando n é pe-

queno, isto é, igual a 100, o limite superior do intervalo de confiança a 95% está acima do considerado um bom ajustamento na literatura, exceto quando o modelo tem dois indicadores em cada fator.

		2 indicadores		3 indicadores		4 indicadores	
		$\lambda = 0,6$	$\lambda = 0,8$	$\lambda = 0,6$	$\lambda = 0,8$	$\lambda = 0,6$	$\lambda = 0,8$
$\phi_{12} = 0,1$	n=100	0,030	0,029	0,072	0,058	0,074	0,062
	n=250	0,020	0,019	0,045	0,037	0,048	0,038
	n=500	0,014	0,013	0,030	0,026	0,034	0,026
$\phi_{12} = 0,25$	n=100	0,033	0,029	0,069	0,058	0,073	0,059
	n=250	0,023	0,018	0,044	0,036	0,045	0,037
	n=500	0,016	0,012	0,030	0,026	0,032	0,026
$\phi_{12} = 0,5$	n=100	0,039	0,026	0,066	0,052	0,068	0,053
	n=250	0,023	0,013	0,040	0,032	0,044	0,033
	n=500	0,016	0,010	0,028	0,022	0,030	0,025

Tabela 4: Limite superior do intervalo de confiança a 95% para o índice SRMR, modelo AFC medido por 2, 3 ou 4 indicadores e dados completos

Analizando a tabela 5, em modelos de AFC com 3 e 4 indicadores medindo cada fator, qualquer que seja a dimensão da amostra e o valor da correlação entre fatores latentes, os valores do índice SRMR aumentam, quando existem omissões nos dados face aos resultados obtidos para os modelos sem omissões. Este aumento é mais acentuado quando as omissões são 25% em cada indicador, do que quando se consideram omissões PMD. Com a existência de omissões nos dados, não é só para amostras de dimensão 100, mas também para amostras de dimensão 250, temos valores acima do aceitável (0,05). De referir ainda que, se obtêm valores piores em modelos de AFC com dois fatores medidos por quatro indicadores, do que quando se tem três indicadores a medir cada fator.

3.3 RMSEA

Analizando a tabela 6 é possível concluir que, os valores obtidos de RMSEA diminuem quando aumenta a dimensão da amostra, bem como, quando aumenta o número de indicadores no modelo, exceto quando se consideram dois indicadores a medir cada fator latente.

		3 indicadores		4 indicadores	
		MCAR	PMD	MCAR	PMD
$\phi_{12} = 0,1$	n=100	0,097	0,083	0,104	0,090
	n=250	0,060	0,051	0,062	0,056
	n=500	0,044	0,035	0,044	0,038
$\phi_{12} = 0,25$	n=100	0,096	0,082	0,103	0,087
	n=250	0,059	0,051	0,062	0,055
	n=500	0,041	0,035	0,044	0,037
$\phi_{12} = 0,5$	n=100	0,091	0,076	0,098	0,083
	n=250	0,054	0,046	0,059	0,051
	n=500	0,039	0,031	0,040	0,036

Tabela 5: Limite superior do intervalo de confiança a 95% para o índice SRMR, modelo AFC medido por 3 ou 4 indicadores, $\lambda = 0,6$ e dados com omissões MCAR e PMD

Este índice não parece ser influenciado pelo valor dos pesos fatoriais, pois os seus resultados mantêm-se sensivelmente os mesmos, quer para valores de $\lambda = 0,6$, quer para valores de $\lambda = 0,8$, exceto no caso mais extremo do modelo de AFC com dois fatores, medidos por dois indicadores, cada. O valor do limite superior de um intervalo a 95% de confiança para este índice apresenta valores superiores ao considerado como um ajustamento aceitável, sendo a dimensão da amostra pequena ($n=100$), exceto quando existem dois indicadores por fator latente. Neste caso, os valores parecem aumentar com o aumento da correlação entre os fatores latentes do modelo. Por outro lado, se o número de indicadores que mede cada fator latente é superior a dois, o valor do RMSEA não parece ser muito afetado pelo valor assumido pela correlação entre fatores latente, uma vez que os valores obtidos são muito próximos.

Quando as amostras geradas apresentam omissões, os valores do índice RMSEA são pouco afetados, em modelos com 3 indicadores em cada fator, quando comparados com os resultados obtidos para modelos com dados completos. Todavia, os valores obtidos indicam um mau ajustamento quando a dimensão da amostra é pequena, qualquer que seja o valor da correlação entre os fatores latentes (ver tabela 7).

Em modelos com 4 indicadores por fator, a existência de omissões

		2 indicadores		3 indicadores		4 indicadores	
		$\lambda = 0,6$	$\lambda = 0,8$	$\lambda = 0,6$	$\lambda = 0,8$	$\lambda = 0,6$	$\lambda = 0,8$
$\phi_{12} = 0,1$	n=100	0,076	0,087	0,105	0,108	0,085	0,086
	n=250	0,053	0,058	0,066	0,066	0,051	0,051
	n=500	0,041	0,049	0,044	0,044	0,034	0,033
$\phi_{12} = 0,25$	n=100	0,098	0,120	0,105	0,108	0,085	0,086
	n=250	0,075	0,092	0,064	0,066	0,051	0,051
	n=500	0,061	0,069	0,044	0,044	0,034	0,033
$\phi_{12} = 0,5$	n=100	0,144	0,158	0,105	0,108	0,083	0,084
	n=250	0,100	0,103	0,063	0,064	0,051	0,051
	n=500	0,072	0,072	0,044	0,044	0,034	0,033

Tabela 6: Limite superior do intervalo de confiança a 95% para o índice RMSEA, modelo AFC medido por 2, 3 ou 4 indicadores e dados completos

provoca um aumento do valor de RMSEA, embora esse efeito seja mais acentuado no caso em que temos omissões MCAR, e em amostras de pequena dimensão (n=100).

		3 indicadores		4 indicadores	
		MCAR	PMD	MCAR	PMD
$\phi_{12} = 0,1$	n=100	0,108	0,102	0,092	0,088
	n=250	0,065	0,064	0,054	0,054
	n=500	0,045	0,044	0,033	0,036
$\phi_{12} = 0,25$	n=100	0,108	0,102	0,092	0,090
	n=250	0,065	0,063	0,054	0,054
	n=500	0,045	0,044	0,034	0,036
$\phi_{12} = 0,5$	n=100	0,109	0,105	0,093	0,090
	n=250	0,065	0,063	0,054	0,054
	n=500	0,044	0,043	0,034	0,036

Tabela 7: Limite superior do intervalo de confiança a 95% para o índice RMSEA, modelo AFC medido por 3 ou 4 indicadores, $\lambda = 0,6$ e dados com omissões MCAR e PMD

4 Conclusões

Com a implementação deste estudo de simulação pretendeu-se aferir qual o efeito nas medidas de qualidade do ajustamento modelo-dados com modelos de AFC, da existência de amostras de pequena dimensão e/ou com omissões de diferentes tipos. Foram considerados três índices de qualidade do ajustamento: χ^2/gl , RMSEA e SRMR.

O índice médio do rácio χ^2/gl apresentou sempre valores abaixo do considerado aceitável na literatura, exceto quando o modelo considerado tinha dois fatores latentes medidos por dois indicadores cada e valores de correlação entre fatores latentes mais elevados. Os índices RMSEA e SRMR apresentaram valores superiores ao considerado aceitável, quando as amostras eram de pequena dimensão e em modelos em que cada fator é medido por mais de dois indicadores. Em modelos de AFC com dois indicadores por fator, apenas o índice RMSEA apresentou valores acima do considerado como um ajustamento aceitável.

Por outro lado, quando existem omissões nos dados os valores obtidos para os três índices de ajustamento pioram, em particular, os índices que se mostraram mais afetados foram o SRMR e o rácio χ^2/gl . Aliás, o índice RMSEA não se mostrou muito afetado pelas omissões PMD. Para obter valores dentro do considerado como aceitável na literatura, a dimensão da amostra em estudo deve ser superior a 250 observações. A existência de 25% de não respostas em todos os indicadores provocaram piores resultados nos índices de ajustamento, do que quando se consideraram omissões planeadas pelo investigador.

De modo geral, podemos concluir que para modelos AFC com e sem omissões, aumentar a dimensão da amostra provoca uma diminuição no valor dos três índices. Maior número de indicadores por fator provoca menores valores dos índices rácio χ^2/gl e RMSEA, enquanto que o índice SRMR tem comportamento contrário. Aumentar o valor da correlação entre fatores latentes não se reflete nos valores obtidos dos índices rácio χ^2/gl e RMSEA, mas os valores obtidos para o índice SRMR diminuem.

De referir ainda que, em algumas situações, quando se considera o modelo com dois indicadores em cada fator latente face ao modelo com três ou quatro indicadores, os valores obtidos para cada um dos índices em análise parecem ser discordantes. Esta situação pode ser consequência de algumas réplicas apresentarem um problema de convergência, neste que é o modelo mais extremo considerado.

Como em qualquer estudo de simulação, também neste existem limitações, que se pretende responder no futuro próximo, tal como, considerar outros desenhos de omissão planeados pelo investigador.

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e Tecnologia, projeto UID/GES/003115/2013.

Referências

- [1] Bollen, K.A. (1989). *Structural Equations with Latent Variables*. Wiley, New York(USA).
- [2] Enders, C.K. (2010). *Applied Missing Data*. The Guilford Press, New York(USA).
- [3] Muthén, B.O., Asparouhov, T. (2002). Using Mplus Monte Carlo simulations in practice: A note on assessing estimation quality and power in latent variable models. *Mplus Web Notes* 1, version 2.
- [4] Muthén, L.K., Muthén, B.O. (1998-2012). *Mplus user's guide, 7th edition*. Muthén e Muthén, Los Angeles(USA).
- [5] Salgueiro, M.F. (2012). *Modelos com Equações Estruturais*. Edições Sociedade Portuguesa de Estatística, Lisboa(Portugal).
- [6] Schumacker, R.E., Lomax, R.G. (2016). *A Beginner's Guide to Structural Equation Modeling, 4th edition*. Routledge, New York(USA).
- [7] Vicente, P.C.R., Salgueiro, M.F. (2016). Efeito de uma variável explicativa na modelação de uma trajetória latente: Estudo de simulação In Cordeiro, C., Ribeiro C., Sousa, C., Gonçalves, M.H., Antunes, N., Silva, M.E. (eds.): *Estatística: Progressos e Aplicações* 283–295, Sociedade Portuguesa de Estatística.

Os sindicatos no feminino: Um ensaio sobre diferentes formas de visualização

Paulo Marques Alves

Instituto Universitário de Lisboa (ISCTE-IUL), DINÂMIA'CET-IUL, Lisboa, Portugal, *paulo.alves@iscte-iul.pt*

Maria do Carmo Botelho

Instituto Universitário de Lisboa (ISCTE-IUL), CIES-IUL, Lisboa, Portugal, *maria.botelho@iscte-iul.pt*

Palavras-chave: Visualização de dados; Análise de *clusters*; Sindicatos; Mulheres; Portugal

Resumo: A representação visual de dados tem uma história secular, mas o debate sobre a visualização de informação ganhou nova dimensão com o conhecimento da perceção visual, com o uso de recursos informáticos e recentemente com a representação de big data. Este estudo propõe e discute representações visuais alternativas, em função dos objetivos e da análise estatística efetuada. Tem como objetivo a recolha, estrutura, análise e representação de dados sobre a presença feminina na direção dos sindicatos da administração pública, entre 2013 e 2016, dados ainda não estudados na perspectiva apresentada. Os resultados revelam uma sub-representação generalizada das mulheres, apresentados em gráficos mais eficientes e segundo boas práticas de visualização.

1 Visualização de dados

A representação visual de dados é uma parte fundamental na análise exploratória e estatística de dados, assim como na comunicação da informação relevante. Uma visualização corresponde a um qualquer tipo de representação visual de informação, destinada a permitir a

comunicação, análise, descoberta, exploração, etc [1].

As representações mais antigas estão associadas a mapas, mas no final do séc. XVIII, William Playfair desempenhou um papel fundamental na criação de diversos gráficos, como por exemplo, a representação da balança comercial de Inglaterra, publicado em 1786, ou posteriormente, em 1801, a criação do primeiro gráfico circular, com o objectivo de representar a relação entre partes e o seu todo [2] [3]. Para Playfair, os gráficos são preferidos às tabelas porque mostram a forma dos dados numa perspetiva comparativa [4]. Tukey [5] considerou que os gráficos são úteis, não só para nos mostrar o que já sabemos, mas também para nos revelar e fazer perceber o que nunca esperávamos ver.

A visualização de dados é uma linguagem e está diretamente relacionada com a captação da imagem pelo olho e o processamento dessa imagem que ocorre no cérebro. Um melhor conhecimento destes processos permite uma optimização das representações visuais a construir. O trabalho de semiologia dos gráficos, desenvolvido por Bertin [6], sistematiza os tipos de codificação visual, para melhor perceção e entendimento da informação. O autor considerou que os dados podem ser codificados em variáveis visuais, como sejam, a posição, o comprimento, intensidade, cor, textura, orientação, e a forma. Estas variáveis foram associadas a quatro propriedades percetivas: a associação, seleção, ordenação e quantificação. Por exemplo, associamos mais facilmente as representações pela forma, cor ou orientação e mais dificilmente pelo comprimento ou pela posição. Baseados na perceção visual e nas tarefas de perceção que as pessoas usam para extrair informação quantitativas dos gráficos, Cleveland e McGill [7], estabeleceram pela primeira vez uma hierarquia de tipos de gráficos, colocando no topo, os de mais fácil e precisa interpretação e o mais confuso e indutor de erro, na base. Desta forma, é possível encontrar no topo gráficos de barras, com codificação dos dados através de comprimento ou posição, com menor perceção comparativa, seguem-se as direções, os ângulos e as áreas. Na base, com menor precisão de análise comparativa, encontram-se os gráficos com volumes, curvas, saturações e intensidades de cor,

recomendados apenas para a percepção de padrões mais genéricos [7], [8], [1], [9].

As técnicas estatísticas são muito importantes na análise de dados, mas para revelar a informação apurada, é essencial ter preocupações com a sua visualização, como por exemplo, estruturar uma tabela de forma apropriada ou saber qual o gráfico mais adequado para o tipo de informação existente, aos objetivos definidos, à audiência a que se destina, entre outros. Não existe um gráfico ideal, mas é possível seguir algumas boas práticas de visualização e facilitar a percepção e leitura da informação mais importante de um gráfico.

Neste artigo pretende-se discutir algumas escolhas de gráficos e alterações de variáveis visuais, para comunicar os resultados encontrados sobre a presença das mulheres nas direções dos sindicatos da administração pública.

2 Trabalho no feminino e a sua relação com os sindicatos

O sindicalismo nasceu andro-centrado, tendo revelado desde o seu início uma atitude sexista em relação ao papel da mulher na sociedade. Ao longo dos tempos, com a incorporação em massa das mulheres no mercado de trabalho, os sindicatos reorientaram as suas estratégias e os seus programas, tendo passado de uma “lógica de exclusão” para uma “lógica de organização”. Bouaffre e Sechi [10] salientam que a tendência para a sub-representação continua, ainda que se verificando diferenças de assinalar: as confederações dos países do sul e do leste da Europa são mais fortemente dominadas pelos homens. A tendência para uma sub-representação mais ou menos intensa ocorre igualmente ao nível das organizações sindicais de primeiro nível, como comprovam [13], para os EUA; [14], para o Reino Unido [15], ou para a Suécia. Para Le Quentrec *et al.* [16], a sub-representação é socialmente construída. A escassez de tempo, dado o trabalho da mulher na esfera privada, é a variável fundamental que

explicará o diferencial de militância entre mulheres e homens. Por forma a melhorar a representação feminina nas confederações sindicais muito tem contribuído a adoção de várias medidas, como sejam a reserva de lugares, as quotas, as comissões de mulheres ou a realização de conferências. Na década passada foi proposta uma nova abordagem impulsionada pelo conceito de *gender mainstreaming*, uma “abordagem integrada da igualdade”, uma nova conceção da igualdade entre homens e mulheres, integrada e permanente, que passa por os sindicatos integrarem este objetivo nas suas práticas (igualdade em matéria de representação nas instâncias dirigentes) e nas suas estratégias. Contudo, diversos estudos têm demonstrado a existência de entraves à implementação do conceito. Garcia [11] propôs que se concedesse atenção às ações já implementadas, como as conferências de mulheres, a criação de comissões de mulheres ou de igualdade, o estabelecimento de quotas, a reserva de lugares ou a garantia de uma representação proporcional e Silvera [12] propôs a integração do tema da igualdade na formação sindical de base, para além do desenvolvimento de ações específicas mais aprofundadas. Apesar dos avanços registados, o reconhecimento e integração das mulheres nos sindicatos e da sua direção, bem como a implementação de uma política de igualdade, continua a encontrar dificuldades e a caracterizar-se pela lentidão.

3 Metodologia

Este estudo pretende contribuir para o estudo da participação das mulheres no movimento sindical em Portugal. Os dados sobre este tema não existem numa única fonte, nem se encontram organizados, sendo assim inovadora a abordagem proposta, quer com a estrutura e organização dos dados, quer na análise estatística multivariada utilizada. Pretende-se também analisar e discutir algumas formas de visualização, o seu contributo para a descoberta de casos atípicos, comparações ou relações entre os dados, de acordo com as melhores práticas de visualização [6], [1], [9]. Pretende-se contribuir para o es-

tudo da participação das mulheres no movimento sindical em Portugal, procurando aferir se no nosso país também se verifica o fenómeno da sub-representação, entendida como uma menor representação feminina nas estruturas dirigentes dos sindicatos por comparação com a proporção de mulheres na população sindicalizável e/ou nos efetivos sindicais. A seleção dos sindicatos da administração pública como campo empírico teve por base três critérios. O primeiro foi a elevada taxa de feminização do emprego que se regista neste sector, 60% no conjunto da administração, no 1º trimestre de 2018 [17]. Os restantes relacionam-se com o sistema sindical existente no sector. Por um lado, ele engloba algumas das estruturas de maior dimensão do país. Por outro lado, é na administração pública (42%), na saúde (44%) e na educação (63%) que se verificam as taxas de feminização das direções sindicais mais elevadas.

Organizaram-se os dados a três níveis: a população feminina sindicalizável num ramo, serviço ou profissão; a constituição das equipas dirigentes; e a liderança da organização. Para apurar a população feminina sindicalizável, recolheu-se a informação estatística oficial disponível, sobre a percentagem de mulheres existente em cada ano e foi determinada uma percentagem média para o período de 2013 a 2016. Para obter os dados referentes à constituição das equipas dirigentes e às lideranças sindicais, realizou-se uma análise documental, com incidência nas fichas biográficas dos dirigentes sindicais, publicadas no Boletim de Trabalho e Emprego e no Jornal Oficial da Região Autónoma da Madeira, na sequência das eleições realizadas (2013-2016). Apurou-se a percentagem de mulheres na direção, na última eleição, que decorreu entre 2013 e 2016. Este intervalo de tempo justifica-se porque as eleições podem ser efectuadas em anos diferentes para diferentes organizações sindicais, sendo que neste intervalo, cada sindicato teve apenas uma eleição.

Os dados recolhidos foram estruturados e armazenados numa base de dados em Excel e em SPSS *Statistics*, com os indicadores apurados para 102 organizações sindicais, uma amostra que corresponde a cerca de um terço do número total de sindicatos existentes em Portugal, e para o emprego público. Em seguida, realizou-se uma análise

estatística para representar e caracterizar os sindicatos em relação à presença das mulheres na direção, atendendo ao ramo onde têm jurisdição. Com o objetivo de determinar o número de *clusters* de sindicatos diferentes aplicou-se uma análise de *clusters* hierárquica, utilizando-se a taxa de feminização do emprego e a percentagem de mulheres na direção como variáveis de segmentação. Em seguida, para a sua classificação, aplicou-se o método k-médias, técnica de análise de *clusters* não hierárquica, de otimização [18].

4 Resultados

Dos 102 sindicatos analisados, 32% têm jurisdição na educação; 19% na saúde; outros 19% nas forças e serviços de segurança; 7% na justiça e 23% foram englobados numa categoria intitulada “outra administração pública”, ou por terem âmbito de atuação noutros ramos da administração pública ou porque lhe são transversais. A maioria (78%) são sindicatos de profissão (49% de profissões não manuais e 29% de profissões científicas e técnicas); 12% têm jurisdição num ramo e 10% num serviço.

O sistema sindical na administração pública apresenta um grau de consistência baixo, consequência do elevado número de organizações que existem, mas também devido ao baixo índice de filiação confederal, ainda que os maiores sindicatos estejam filiados nas estruturas de topo do movimento sindical português. Verifica-se que 68% dos sindicatos não são filiados confederalmente; 18% são filiados na UGT e 14% na CGTP-IN.

Entre os sindicatos da administração pública é muito baixa a proporção de mulheres nos cargos de liderança. A percentagem média é de apenas 17%, apresentando os sindicatos da educação e da justiça valores mais elevados (24% e 29% respetivamente).

Ao analisar as equipas dirigentes, verifica-se que as taxas de feminização das direções são, em geral, baixas. Os sindicatos associados à educação, surgem como exceção, com uma elevada ou mesmo muito

elevada participação feminina ($>60\%$), em 63% dos casos, e, em menor escala, na justiça (média ou elevada participação – 41%-80%). Saliente-se que existem nove sindicatos que não apresentam qualquer mulher na direção, sendo que oito pertencem às forças e serviços de segurança. Apenas uma organização evidencia uma taxa de feminização da direção de 100% (SIMAC⁸) e outro, uma taxa superior a 90% (SPCL⁹).

A sub-representação das mulheres nas direções sindicais da administração pública é generalizada. Contudo, é necessário tomar em consideração que existem ramos onde a feminização do emprego é baixa, como é o caso das forças e serviços de segurança, sendo expectável uma fraca presença das mulheres nas estruturas de decisão dos respetivos sindicatos. Procedemos ao confronto da percentagem de mulheres nas direções com a taxa de feminização do emprego, ou seja, a população feminina potencialmente sindicalizável em cada ramo. A Figura 1 pretende representar a comparação dos dois indicadores, com uma opção gráfica obtida com o Excel, de forma automática. Este gráfico é classificado como uma opção que permite um maior rigor comparativo [1], com o uso de duas das variáveis visuais, a cor e a dimensão [6]. Contudo, não facilita a imediata perceção da comparação que é pretendida. Para além das opções automáticas das cores não serem as mais adequadas, as linhas auxiliares surgem em excesso, tal como as divisões na escala do eixo vertical. Para corrigir estas dificuldades e salientar a comparação, foram construídas duas alternativas ao gráfico inicial.

A Figura 2 mostra as duas representações construídas em Excel, onde são percecionadas de forma mais imediata as diferenças entre a participação das mulheres na direção dos sindicatos e a população potencialmente sindicalizável, porque se encontram as comparações na vertical e não lado a lado, usando a cor e a forma para os dois indicadores diferentes, continuando sobre o mesmo eixo. A versão B mostra como meta a taxa de feminização, revelando que seria pelo

⁸Sindicato Nacional de Massagistas de Recuperação e Cinesioterapeutas.

⁹Sindicato dos Professoras nas Comunidades Lusíadas.

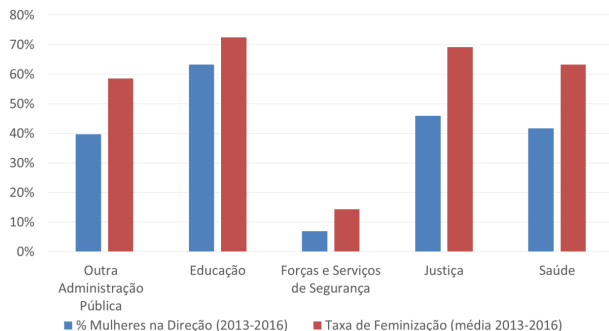


Figura 1: Percentagem de mulheres nas direções dos sindicatos da administração pública e taxa de feminização do emprego (%), por ramo, em Portugal (2013-2016)

Fonte: Cálculos próprios a partir do BTE e JORAM

menos razoável atingir este valor na direção dos sindicatos em cada ramo. Acresce ainda uma classificação em fundo, em três níveis de participação feminina (baixa, média e elevada). Este gráfico, também conhecido como *bullet chart* [9], permite um maior nível de análise e de classificação associados à imagem. Faz uso da forma e da cor para a comparação entre o observado e o objetivo. Utiliza também a cor, com escala de intensidade, para evidenciar a classificação ordinal do nível de participação das mulheres em cada ramo, deixando com mais intensidade a tradução da situação menos favorável (baixa representação das mulheres).

No que concerne à interpretação, pela observação da Figura 2 verifica-se, por um lado, que as forças e serviços de segurança registam uma diminuta presença de mulheres nas direções, mas apresentam um diferencial de apenas sete pontos percentuais (p.p.) face à taxa de feminização destas forças. Por outro lado, o ramo da educação é o que apresenta uma maior presença de mulheres e é também o que regista uma sua maior proporção nas direções sindicais, se bem que se verifique um diferencial de dez p.p.. Os sindicatos da saúde e da justiça são os que manifestam um maior afastamento, com um

diferencial de 23 p.p. e 21 p.p. respetivamente.

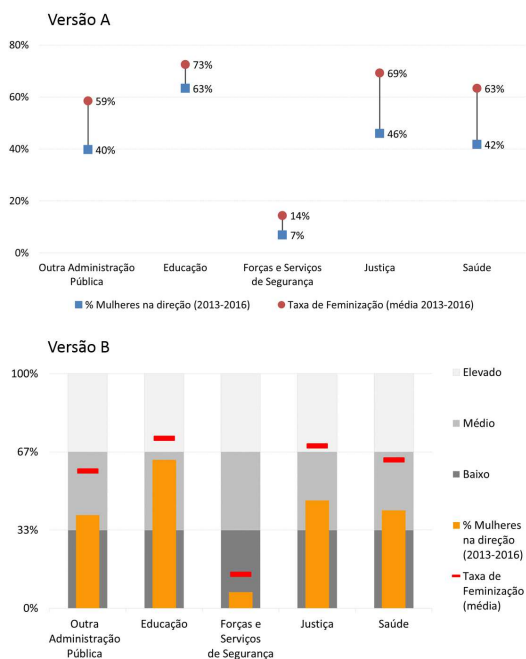


Figura 2: Diferença entre a percentagem de mulheres nas direções dos sindicatos da administração pública e a taxa de feminização do emprego (%), por ramo, em Portugal (2013-2016)

Fonte: Cálculos próprios a partir do BTE e JORAM

Pode ainda ser referido que apenas a educação e a justiça poderiam vir a atingir uma elevada participação das mulheres na direção, enquanto nas forças e serviços de segurança a presença é muito baixa e mesmo que atingisse uma percentagem idêntica à presença feminina no ramo, esta seria sempre classificada como baixa, quando comparada com os restantes ramos. De acordo com os dois indicadores mencionados, percentagem de mulheres na direção e taxa de

feminização do emprego, procedeu-se a uma segmentação dos sindicatos. Numa primeira etapa, para mostrar a forte relação positiva entre estes dois indicadores ($R=0,809$), construiu-se um gráfico de quadrantes em Excel, utilizando a média global de cada indicador como referencial para os eixos (Figura 3). Esta representação visual surge adequada para perceber a relação entre dois indicadores, mas permite também, numa análise exploratória, a descoberta de casos atípicos situados nos quadrantes pares, passíveis de um posterior aprofundamento da análise [1].

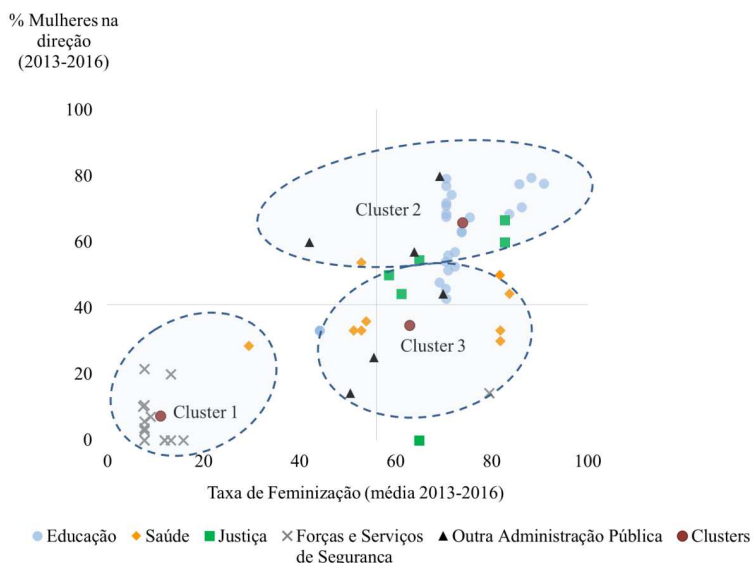


Figura 3: Tipologia dos sindicatos da administração pública em Portugal (2013-2016) (n=61)

Fonte: Elaboração própria

Assim, no segundo quadrante encontram-se o SMZS¹⁰, no ramo da

¹⁰Sindicato dos Médicos da Zona Sul.

saúde, e o SINSEF¹¹, que integrámos na categoria “outra administração pública”, são sindicatos com uma taxa de feminização inferior à média do conjunto e com uma percentagem de mulheres na direcção superior à média. No quarto quadrante encontram-se o SPTAAP¹², o SFJ¹³ e dois sindicatos do ramo da saúde, o SIPE¹⁴ e o SE¹⁵, sindicatos com uma taxa de feminização superior à média do conjunto, mas onde a presença de mulheres na direcção é inferior à respetiva média.

Numa segunda etapa de trabalho com os dados, foi aplicada uma análise de *clusters* hierárquica. Como variáveis de segmentação foram utilizadas a percentagem de feminização e a percentagem de mulheres na direcção, para o período de tempo considerado. Utilizaram-se os métodos de agregação de *Ward* e do vizinho mais afastado, ambos com o quadrado da distância Euclideana para o cálculo das distâncias. Depois de analisados os dendogramas resultantes e os coeficientes de fusão encontrados nas etapas de agregação com cada um dos métodos, ambos sugeriam a formação de três *clusters* distintos de sindicatos. Em seguida a formação dos *clusters* foi efectuada com recurso ao método não hierárquico k-médias, peritindo assim a optimização do agrupamento realizado. Na Figura 3) foram projectados os centróides finais dos três clusters, encontrados com o recurso ao K-médias e foi desenhada uma linha a tracejado, orientadora para a sua caracterização.

O primeiro *cluster* ($n_1=15$) é composto por sindicatos com jurisdição em ramos com uma baixa taxa de feminização do emprego e que apresentam em consequência uma muito fraca representação das mulheres nas direcções. Este *cluster* é constituído quase exclusivamente por organizações das forças e serviços de segurança (93%). O segundo *cluster* ($n_2=27$) é integrado por sindicatos que atuam em ramos com uma elevada taxa de feminização do emprego, re-

¹¹Sindicato dos Funcionários Serviço de Estrangeiros Fronteiras.

¹²Sindicato Pessoal Técnico Apoio Atividade Policial da PSP.

¹³Sindicato dos Funcionários Judiciais.

¹⁴Sindicato Independente Profissionais de Enfermagem.

¹⁵Sindicato dos Enfermeiros.

gistando uma presença de mulheres nas direções sindicais acima da média, sendo formado maioritariamente por estruturas do ramo da educação (74%). O terceiro *cluster* ($n_3=19$) engloba sindicatos com uma taxa de feminização semelhante ou superior à média, com as organizações a evidenciarem uma percentagem de mulheres na direcção muito inferior ao seu peso na população sindicalizável. Compõem-no sobretudo sindicatos da saúde (37%) e da educação (26%).

Para uma amostra de 63 sindicatos foi possível apurar a presença de mulheres na direcção entre 2003 e 2006, ou seja, 10 anos antes. Com estes dados, procedeu-se a uma análise comparativa temporal. Verificou-se a existência de um retrocesso das taxas de feminização das direcções num conjunto elevado de organizações (27), ligados ao ramo da educação. Os restantes ramos revelaram um ligeiro aumento.

5 Conclusão

A presente investigação pretende contribuir para o estudo da representação das mulheres nas estruturas sindicais, particularmente nas suas direcções. Apresenta a inovação da criação de uma base organizada e classificada de dados sobre os sindicatos da administração pública, assim como a análise multivariada realizada. Paralelamente, também foram discutidas as opções de visualização, adotando as melhores práticas neste domínio. Apresentam-se possibilidades de modificação das representações gráficas “tradicionais”, com recurso ao Excel. A escolha da visualização de dados mais adequada relaciona-se com a mensagem que se pretende comunicar, com a análise que se pretende fazer, podem ser comparações, relações, classificações ou deteção de padrões. Sugere-se a construção de mais do que uma proposta de gráfico, arranjadas as suas componentes, e decidir sobre qual a forma que facilita a extração de conhecimento a partir da imagem [1].

Tradicional ou não, recomenda-se a simplicidade na apresentação de valores, linhas, formas ou cores, de acordo com a referência atual de “*keep it simple*”. Nas modificações a efectuar não devem alterar

mais de duas variáveis visuais [6], [1].

Relativamente aos resultados encontrados sobre a militância sindical no feminino, este estudo permitiu concluir que mesmo em ramos e/ou profissões altamente feminizados, isso não é garantia de uma adequada representação das mulheres. Apesar dos vários padrões que descortinámos, isso é notório para o conjunto dos sindicatos da administração pública, em particular no caso dos sindicatos dos profissionais de enfermagem, onde encontramos diferenciais superiores a 40 p.p. entre a proporção de mulheres na população sindicalizável e a sua proporção nas instâncias de decisão. Uma não adequada representação das mulheres tem consequências nefastas para o sindicalismo. Por um lado, se ele é menos inclusivo então é menos representativo. Por outro, como alertam vários autores [19], [14], se as lideranças sindicais não representam de forma proporcional os efetivos, então o carácter democrático das organizações é fortemente restringido. Se bem que se mantenham insuficiências, a situação evoluiu bastante por comparação com as décadas de 70 e 80 e, mesmo, de 90. Muito há a fazer, mas os sindicatos saberão encontrar e aplicar as políticas que permitirão acabar com a sub-representação das mulheres nos seus órgãos dirigentes.

Referências

- [1] Cairo, A. (2016). *The Truthful Art: Data, Charts, and Maps for Communication*. New Riders.
- [2] Friendly, M. (2009). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. Disponível em <http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf> (Acedido em 10 de setembro de 2015).
- [3] Alexandrino da Silva, A. (2006). *Gráficos e mapas. Representação de informação estatística*. Lidel, Lisboa.
- [4] Tufte, E.R. (2001). *The Visual Display of Quantitative Information*. Graphics Press, Connecticut.
- [5] Tukey, J.W. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company.

- [6] Bertin, J. (1967). *Semiologie Graphique*. Gauthier-Villars, Paris.
- [7] Cleveland, w.S., McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. American Statistical Association*, 79(387), 531–554.
- [8] Cairo, A. (2013). *The Functional Art. An Introduction to Information Graphics and Visualization*. New Riders.
- [9] Evergreen, S. D. H. (2017). *Effective Data Visualization. The Right Chart for the Right Data*. SAGE, Thousand Oaks.
- [10] Bouaffre, A., Sechi, C. (2014). *Tendances de L’affiliation Féminine au sein des Confédérations Syndicales Nationales*. ETUI, Bruxelles.
- [11] Garcia, A. (2003). *Femmes dans les Syndicats. Une Nouvelle Donne*. ETUI, Bruxelles.
- [12] Silvera, R. (2006). Le défi de l’approche intégrée de l’égalité pour le syndicalisme en Europe. *La Revue de l’IRES* 50, 137–172.
- [13] Cobble, D., Michal, M. (2002). On the edge of equality? Working women and the US labor movement. In Colgan, F., Ledwith, S. (eds.): *Gender, Diversity and Trade Unions. International Perspectives*. Routledge, Londres.
- [14] Colgan, F., Ledwith, S. (1996). Sisters organizing: women and their trade unions. In Ledwith, S., Colgan F. (eds.): *Women in Organizations: Challenging Gender Politics*. Macmillan, Londres.
- [15] Mahon, R. (2002). Sweden’s LO: learning to embrace the differences within? In Colgan, F., Ledwith, S. (eds.): *Gender, Diversity and Trade Unions. International Perspectives*. Routledge, Londres.
- [16] Le Quentrec, Y., Rieu, A., Lapeyre, N. (1999). *Femmes dans la prise de décision syndicale: pour quels changements?*. Comunicação apresentada às Journées d’Études Doctorales Interdisciplinaires sur le Syndicalisme. Paris.
- [17] DGAEP (2018). *Síntese Estatística do Emprego Público - 3º Trimestre de 2018*. DGAEP, Lisboa.
- [18] Hair, J.F., Black W.C., Babin, B.J., Anderson R.E. (2010). *Multivariate data analysis: a global perspective*. 7th ed, Upper Saddle River, N.J: Pearson Education.
- [19] Cockburn, C. (1991). *In the Way of Women: Men’s Resistance to Sex Equality in Organizations*. Cornell University Press, Ithaca.

Distribuição limite conjunta da soma e do máximo de variáveis inteiras

Sandra Dias

CEMAT, Departamento de Matemática, Universidade de Trás-os-Montes e Alto Douro, *sdias@utad.pt*

Maria da Graça Temido

CMUC, Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade de Coimbra, *mgtm@mat.uc.pt*

Palavras-chave: Teoria de valores extremos; Classe de Anderson; Lei max-semiestáveis; Leis semiestáveis.

Resumo: Neste trabalho consideramos sucessões fortemente estacionárias, $\{X_n\}$, cujas margens são variáveis aleatórias inteiras positivas pertencentes à classe de Anderson ([3]). É estudada a distribuição assintótica de (S_{k_n}, M_{k_n}) , onde $S_{k_n} = X_1 + \dots + X_{k_n}$, $M_{k_n} = \max\{X_1, \dots, X_{k_n}\}$ e $\{k_n\}$ é uma sucessão de inteiros positivos não decrescente tal que $k_{n+1}/k_n \rightarrow r > 1$, $n \rightarrow +\infty$. Considerando este crescimento geométrico da dimensão amostral, é estabelecido um limite não degenerado para $P(S_{k_n} \leq a_n x + b_n, M_{k_n} \leq x + d_n)$, o qual expressa a independência assintótica destas estatísticas. O resultado é ilustrado com uma classe de modelos Ψ -INAR(1).

1 Introdução

Nas últimas décadas, tem-se verificado um interesse crescente pela modelação de dados ambientais, como a temperatura, a velocidade do vento, os níveis de água, a quantidade de poluentes, onde a ocorrência de valores elevados pode influenciar significativamente a sua média. Se, para amostras de dimensão moderada, tal influência é notória, para grandes dimensões amostrais, a dependência (assintótica)

entre média e máximo nem sempre se verifica. Assim, há que estudar a distribuição assintótica conjunta da soma $S_n = X_1 + \dots + X_n$ e do máximo $M_n = \max\{X_1, \dots, X_n\}$ de uma sucessão de variáveis aleatórias (v.a.'s) $\{X_n\}$.

Um dos trabalhos pioneiros deve-se a Chow e Teugels [4], que considerando as v.a.'s $\{X_n\}$ independentes e identicamente distribuídas (i.i.d.), mostraram que se $(a_n^{-1}S_n + nb_n, c_n^{-1}(M_n - d_n))$, com $a_n, c_n > 0$, b_n e d_n constantes reais, converge em distribuição para (U, V) então U e V são independentes, exceto quando o índice de estabilidade $\alpha < 2$ e V tem distribuição de Fréchet. Neste caso, os autores estabelecem a função característica de (U, V) no limite através de

$$E\left(e^{it(a_n^{-1}S_n + nb_n)} \mathbb{I}_{\{M_n \leq v_n\}}\right) \longrightarrow w_\alpha(t, p) \exp\left[\int_v^{+\infty} e^{itkw} dw^{-\alpha}\right],$$

onde $v_n = c_n v + d_n$, $w_\alpha(t, p)$ é a função característica limite de $a_n^{-1}S_n + nb_n$ e k é uma constante. Com a motivação de encontrar um modelo adequado à distribuição conjunta da velocidade média e máxima do vento, entre outras, a literatura prosseguiu com os resultados de Anderson e Turkman [2], relativos também à distribuição limite de (S_n, M_n) , considerando todavia que $\{X_n\}$ é uma sucessão fortemente estacionária (estacionária) sob condições de dependência adequadas.

Como em muitas áreas, como filas de espera, epidemiologia e meteorologia, surgem séries temporais de valores inteiros, há que estabelecer o comportamento limite de (S_n, M_n) , tendo por suporte os modelos de variáveis inteiras e as técnicas subjacentes.

Em McCormick e Sun [15] é estudado o comportamento assintótico deste par aleatório, assumindo que a sucessão subjacente $\{X_n\}$ é estacionária com margens pertencentes à classe de Anderson ([3]), isto é, à classe das funções de distribuição (f.d.'s) F com extremo superior do suporte, w_F , infinito e que verificam

$$\frac{1 - F(n-1)}{1 - F(n)} \rightarrow r > 1, n \rightarrow +\infty, \quad (1)$$

o que denotamos por $F \in \mathcal{C}_A(r)$. Sob normalização linear, são obtidos os limites inferior e superior (diferentes) para tal distribuição.

Mantendo a plataforma probabilista em que as variáveis da sucessão estacionária $\{X_n\}$ são inteiras positivas e pertencem à classe de Anderson, neste trabalho é estudada a distribuição assintótica de (S_{k_n}, M_{k_n}) , onde $\{k_n\}$ é uma sucessão de inteiros positivos não decrescente tal que

$$k_{n+1}/k_n \rightarrow r > 1, n \rightarrow +\infty. \quad (2)$$

Na Secção 2, considerando que as margens de $\{X_n\}$ são i.i.d., são apresentadas as distribuições limite de $P(S_{k_n} \leq a_n x + b_n)$ e $P(M_{k_n} \leq x + d_n)$, que caso existam são semiestáveis e max-semiestáveis, respectivamente. Na Secção 3 é dedicada a alguns resultados sobre sucessões estacionárias e f.d.'s pertencentes à classe de Anderson, no contexto max-semiestável. Na Secção 4 estabelece-se um limite não degenerado para $P(S_{k_n} \leq a_n x + b_n, M_{k_n} \leq x + d_n)$, o qual explicita a independência assintótica destas estatísticas. A Secção 5 encerra uma aplicação do resultado anterior a uma classe de modelos Ψ -INAR(1).

2 Leis semiestáveis e max-semiestáveis

Apresentamos nesta secção alguns resultados sobre as distribuições limite da soma e do máximo, S_{k_n} e M_{k_n} , de k_n v.a.'s i.i.d. onde $\{k_n\}$ é uma sucessão de inteiros que satisfaz (2). Para o que se expõe nesta secção as v.a.'s não são necessariamente inteiras.

Resultados relativos à distribuição limite de S_{k_n} surgem nos trabalhos de Kruglov [12] e Grinevich e Khokhlov [8], onde se prova que a distribuição limite da sucessão $\{a_n^{-1} S_{k_n} + b_n\}$, com $a_n > 0$ e b_n reais, é uma função distribuição (f.d.) semiestável G . Nesse caso, dizemos que a f.d. F pertence ao domínio de atração de uma f.d. semiestável G_α , e escrevemos $F \in \mathcal{D}(G_\alpha)$. Uma f.d. semiestável G_α ou é Gaussiana ou então a sua função característica, φ , admite a

representação

$$\log \varphi(t) = ict + \int_{-\infty}^{+\infty} \left(e^{itx} - 1 - \frac{itx}{1+x^2} \right) dH(x), \quad (3)$$

com $c \in \mathbb{R}$ e função espectral

$$H(x) = \begin{cases} (-x)^{-\alpha} \theta_1(\log(-x)) & \text{se } x < 0 \\ -x^{-\alpha} \theta_2(\log x) & \text{se } x > 0 \end{cases}, \quad (4)$$

onde $0 < \alpha < 2$, θ_i são funções periódicas com período comum e tais que, para todo o x e para todo o $h \geq 0$, $e^{\alpha h} \theta_i(x-h) - e^{-\alpha h} \theta_i(x+h) \geq 0$, $0 \leq c_i \leq \theta_i(x) \leq d_i < +\infty$ e $c_1 + c_2 > 0$. A $\alpha \in]0, 2]$ chamamos de índice de semiestabilidade.

Notamos que, se em (2) substituirmos r por 1, então a função espectral H é definida por $H(x) = (-x)^{-\alpha} \theta_1$ se $x < 0$ e $H(x) = -x^{-\alpha} \theta_2$ se $x > 0$, onde θ_1 e θ_2 são constantes não simultaneamente nulas, e assim G_α representa uma f.d. estável. A classe de distribuições max-semiestáveis foi introduzida em Grinevich [6, 7]. Nesses trabalhos provou-se que, se $\{k_n\}$ é uma sucessão de inteiros satisfazendo (2) então a distribuição limite da sucessão $\{(M_{k_n} - d_n)/c_n\}$, com $c_n > 0$ e d_n constantes reais, é uma f.d. max-semiestável $G_{\beta, \nu}$. Dizemos assim que a f.d. F pertence ao domínio de atração da f.d. max-semiestável $G_{\beta, \nu}$ e escrevemos $F \in \mathcal{D}(G_{\beta, \nu})$. Uma f.d. max-semiestável $G_{\beta, \nu}$ é do tipo de uma das f.d.'s caracterizadas por

$$\begin{aligned} \Phi_{\beta, \nu}(x) &= \exp(-x^{-\beta} \nu(\log(x))), \quad x > 0 \\ \Psi_{\beta, \nu}(x) &= \exp((-x)^{\beta} \nu(\log(-x))), \quad x < 0 \\ \Lambda_{\nu}(x) &= \exp(-e^{-x} \nu(x)), \quad x \in \mathbb{R}, \end{aligned}$$

onde $\nu(x)$ é uma função positiva, limitada e periódica com período $\log r$. Também aqui observamos que, se em (2) substituirmos r por 1, então o limite em distribuição de M_{k_n} , sob normalização linear, encontra-se na classe das bem conhecidas f.d.'s max-estáveis.

3 Sucessões estacionárias, classe de Anderson e max-semiestabilidade

Seja $\{X_n\}$ uma sucessão estacionária com f.d. marginal F . Recorde-mos algumas condições de dependência das sucessões estacionárias, em particular em contexto de max-semiestabilidade.

Definição 3.1 *A sucessão $\{X_n\}$ verifica a condição de mistura forte se, para quaisquer acontecimentos A e B pertencentes à σ -álgebra gerada por $\{\dots, X_{t-1}, X_t\}$ e $\{X_{t+n}, X_{t+n+1}, \dots\}$, se tem*

$$|P(A \cap B) - P(A)P(B)| \leq \alpha(n), \quad \forall n \in \mathbb{N}, \quad \forall t \in \mathbb{N},$$

para alguma sucessão não decrescente $\alpha(n)$, com $\lim_{n \rightarrow +\infty} \alpha(n) = 0$.

As condições de independência assintótica $D_{k_n}(u_n)$ e de dependência local $D'_{k_n}(u_n)$, que são adaptações das condições $D(u_n)$ e $D'(u_n)$ de Leadbetter *et al.* [13] ao contexto de max-semiestabilidade, são introduzidas em Temido e Canto e Castro [17] e Hall e Temido [10], respetivamente.

Definição 3.2 *Seja $\{k_n\}$ uma sucessão crescente de inteiros positivos e $\{u_n\}$ uma sucessão de reais. A sucessão $\{X_n\}$ verifica a condição $D_{k_n}(u_n)$ se, para quaisquer inteiros $1 \leq i_1 < \dots < i_p < j_1 < \dots < j_q \leq k_n$, com $j_1 - i_p > \ell_n$ e $A_j := \{X_j \leq u_n\}$, se tem*

$$|P(\cap_{s=1}^p A_{i_s}, \cap_{m=1}^q A_{j_m}) - P(\cap_{s=1}^p A_{i_s})P(\cap_{m=1}^q A_{j_m})| \leq \alpha_{n, \ell_n},$$

onde $\lim_{n \rightarrow +\infty} \alpha_{n, \ell_n} = 0$, para alguma sucessão ℓ_n a satisfazer $\lim_{n \rightarrow +\infty} \ell_n / k_n = 0$.

Temido e Canto e Castro [17] provam que, sob $D_{k_n}(u_n)$, a distribuição limite do máximo M_{k_n} , se existir, é max-semiestável.

Definição 3.3 *Seja $\{k_n\}$ uma sucessão crescente de inteiros positivos e $\{u_n\}$ uma sucessão de reais. A sucessão $\{X_n\}$ satisfaz a*

condição $D'_{k_n}(u_n)$ se existir uma sucessão de inteiros positivos $\{s_n\}$ tal que $\lim_{n \rightarrow +\infty} k_n/s_n = +\infty$, $\lim_{n \rightarrow +\infty} s_n \alpha_{n,l_n} = 0$ e

$$\lim_{n \rightarrow +\infty} k_n \sum_{j=2}^{[k_n/s_n]} P(X_1 > u_n, X_j > u_n) = 0.$$

A noção de índice extremal é também estendida ao contexto de max-semiestabilidade. Para tal, defina-se o conjunto

$$\Gamma(F, k_n) = \{\tau > 0 : \exists u_n(\tau, k_n) : \lim_{n \rightarrow +\infty} k_n(1 - F(u_n(\tau, k_n))) = \tau\}.$$

Note-se que se F é uma f.d. discreta e (2) ocorre, então $\Gamma(F, k_n)$ é um conjunto pelo menos infinito numerável. Define-se que a sucessão $\{X_n\}$ tem índice extremal θ , se existe uma sucessão de inteiros não-decrescente satisfazendo (2) tal que, para todo o $\tau \in \Gamma(F, k_n)$, se tem

$$\lim_{n \rightarrow +\infty} P(M_{k_n} \leq u_n(\tau, k_n)) = e^{-\theta\tau}, \theta \in]0, 1].$$

Como se espera, sob $D_{k_n}(u_n)$ e $D'_{k_n}(u_n)$, os máximos M_{k_n} e F^{k_n} têm o mesmo limite em distribuição, do que decorre $\theta = 1$.

Relativamente às f.d.'s pertencentes à classe $\mathcal{C}_A(r)$ e a sua relação com as f.d.'s max-semiestáveis apresentamos alguns resultados.

Começamos por recordar que as f.d.'s pertencentes a $\mathcal{C}_A(r)$, como a Binomial Negativa, não pertencem a qualquer domínio de atração max-estável. Porém, Anderson [3] provou que, uma f.d. F , com $w_F = \infty$, verifica (1) se e só se existe $\{d_n\}$ tal que, para $u_n = x + d_n$, $\exp(-r^{-x+1}) \leq \liminf_{n \rightarrow +\infty} F^n(u_n)$ e $\limsup_{n \rightarrow +\infty} F^n(u_n) \leq \exp(-r^{-x})$.

Na procura de uma f.d. limite não degenerada, Temido [18] provou que, para qualquer $F \in \mathcal{C}_A(r)$, o limite de F^{k_n} , sob normalização linear, é uma f.d. max-semiestável discreta, designada Gumbel discreta e definida por $G(x) = \exp(-r^{-[x]})$, $x \in \mathbb{R}$. Concretamente, para uma f.d. F com $w_F = +\infty$, existe uma sucessão de inteiros não-decrescente $\{k_n\}$ a satisfazer (2) e uma sucessão $\{d_n\}$ tal que

$$\lim_{n \rightarrow +\infty} F^{k_n}(x + d_n) = \exp(-r^{-[x]}), x \in \mathbb{R},$$

se e só se $F \in \mathcal{C}_A(r)$. Em conclusão, se $\{X_n\}$ é uma sucessão estacionária, com f.d. marginal $F \in \mathcal{C}_A(r)$, e que satisfaz as condições $D_{k_n}(u_n)$ e $D'_{k_n}(u_n)$, então a f.d. limite de $M_{k_n} - d_n$ é também a Gumbel discreta.

4 Distribuição limite de (S_{k_n}, M_{k_n})

Antes de apresentar o resultado principal deste trabalho, há que definir a nova condição, $D'_{k_n}(\sqrt{k_n}, u_n)$, que é uma adaptação da condição $D'(a_n, u_n)$ de Anderson e Turkman [2].

Definição 4.1 *Seja $u_n = u_n(\tau, k_n)$ para qualquer $\tau \in \Gamma(F, k_n)$. A sucessão $\{X_n\}$ satisfaz a condição $D'_{k_n}(\sqrt{k_n}, u_n)$ se*

$$\lim_{s \rightarrow +\infty} \limsup_{n \rightarrow +\infty} s \sum_{j=1}^{r_n} E \left[\left| \exp \left(\frac{it}{\sigma \sqrt{k_n}} \sum_{\substack{l=1 \\ l \neq j}}^{r_n} \tilde{X}_l \right) - 1 \right| \mathbb{1}_{\{X_j > u_n\}} \right] = 0,$$

onde $\sigma^2 = \lim_{n \rightarrow +\infty} \frac{1}{k_n} V \left(\sum_{i=1}^{k_n} X_i \right)$, $\tilde{X}_l = X_l - E(X_l)$ e $r_n = \left[\frac{k_n}{s} \right]$.

O teorema seguinte é o resultado principal deste trabalho, no qual se estabelece um limite não degenerado para (S_{k_n}, M_{k_n}) , evidenciando, como já referimos, a independência assintótica entre estas duas estatísticas.

Teorema 4.2 *Seja $\{k_n\}$ uma sucessão de inteiros positivos não decrescente a satisfazer (2) e $\{X_n\}$ uma sucessão estacionária que verifica a condição de mistura forte e com f.d. marginal, F , pertencente a $\mathcal{C}_A(r)$. Se existirem constantes reais d_n , $a_n > 0$ e b_n tais que $\{X_n\}$ verifica as condições $D'_{k_n}(u_n)$ e $D'_{k_n}(\sqrt{k_n}, u_n)$, com $u_n = y + d_n$, e*

$$a_n^{-1}(S_{k_n} - k_n b_n) \xrightarrow{d} N(0, 1), \quad n \rightarrow +\infty,$$

então,

$$\lim_{n \rightarrow +\infty} P\left(\frac{S_{k_n} - k_n b_n}{a_n} \leq x, M_{k_n} \leq u_n\right) = F_{N(0,1)}(x) \exp(-r^{-[y]}).$$

Dem.: Uma vez que a condição de mistura forte implica a condição $D_{k_n}(u_n)$, com a condição $D'_{k_n}(u_n)$ garante-se a existência de índice extremal igual a 1. Por outro lado, para a classe de Anderson, tem-se a existência de uma sucessão real $\{d_n\}$ tal que $F^{k_n}(x + d_n) \rightarrow \exp(-r^{-[x]})$, $n \rightarrow +\infty$. Assim, a sucessão aleatória $M_{k_n} - d_n$ é atraída em distribuição para o mesmo limite.

Em Temido e Canto e Castro [17] é introduzida uma extensão da divisão em blocos de Loynes, adaptada ao contexto de max-semiestabilidade, com a qual foi possível estabelecer

$$\lim_{s \rightarrow +\infty} \limsup_{n \rightarrow +\infty} s |P(M_{r_n} \leq u_n) - G^{1/s}(x)| = 0. \quad (5)$$

Com os mesmos argumentos e seguindo Anderson e Turkman [2], com $\tilde{S}_{k_n} := S_{k_n} - k_n b_n$ e $\tilde{S}_{r_n} := S_{r_n} - k_n b_n$, também no presente contexto se prova que

$$\lim_{s \rightarrow +\infty} \limsup_{n \rightarrow +\infty} s \left| E \left(e^{it \frac{\tilde{S}_{k_n}}{a_n}} \mathbb{I}_{\{M_{k_n} \leq u_n\}} \right) - E^s \left(e^{it \frac{\tilde{S}_{r_n}}{a_n}} \mathbb{I}_{\{M_{r_n} \leq u_n\}} \right) \right| = 0,$$

bem como

$$\lim_{s \rightarrow +\infty} s \limsup_{n \rightarrow +\infty} E \left[\left(e^{it a_n^{-1} \tilde{S}_{r_n}} - 1 \right) \mathbb{I}_{\{M_{r_n} > u_n\}} \right] = 0.$$

Na prova de este último limite é usada a versão discreta do Lema 2.2 de Anderson e Turkman [2]. A saber, aplicando a soma por partes obtemos

$$\begin{aligned} E(|X_0| \mathbb{I}_{\{X > u_n\}}) &= \sum_{i=[u_n]+1}^{+\infty} iP(X_0 = i) = P(X_0 > u_n)([u_n] + 1) \\ &+ \sum_{i=[u_n]+1}^{+\infty} P(X_0 > i) = O\left(\frac{u_n}{k_n}\right) + o_n(1), \end{aligned}$$

uma vez que $P(X_0 > i + u_n)/P(X_0 > u_n) \longrightarrow r^{-i}$, $n \longrightarrow +\infty$.

Da normalidade assintótica de $a_n^{-1}\tilde{S}_{k_n}$ decorre

$$\lim_{s \rightarrow +\infty} \limsup_{n \rightarrow +\infty} s \left| E \left(e^{ita_n^{-1}\tilde{S}_{r_n}} \right) - e^{-\frac{t^2}{2s}} \right| = 0.$$

Por outro lado

$$\begin{aligned} & \left| E^s \left(e^{ita_n^{-1}\tilde{S}_{r_n}} \mathbb{I}_{\{M_{r_n} \leq u_n\}} \right) - \left(G^{1/s}(x) + e^{-\frac{t^2}{2s}} - 1 \right)^s \right| \\ & \leq s \left| E \left[\left(e^{ita_n^{-1}\tilde{S}_{r_n}} + 1 - 1 \right) (1 - \mathbb{I}_{\{M_{r_n} > u_n\}}) \right] - G^{1/s}(x) - e^{-\frac{t^2}{2s}} + 1 \right| \\ & \leq s \left| E \left[\left(e^{ita_n^{-1}\tilde{S}_{r_n}} - 1 \right) \mathbb{I}_{\{M_{r_n} > u_n\}} \right] \right| + s \left| P(M_{r_n} \leq u_n) - G^{1/s}(x) \right| \\ & \quad + s \left| E \left(e^{ita_n^{-1}\tilde{S}_{r_n}} \right) - e^{-\frac{t^2}{2s}} \right|. \end{aligned}$$

Assim, atendendo aos limites já estabelecidos, obtemos

$$\begin{aligned} & \lim_{s \rightarrow +\infty} \lim_{n \rightarrow +\infty} E^s \left(e^{ita_n^{-1}\tilde{S}_{k_n}} \mathbb{I}_{\{M_{k_n} \leq u_n\}} \right) = \\ & = \lim_{s \rightarrow +\infty} \lim_{n \rightarrow +\infty} \left(G^{1/s}(x) + e^{-\frac{t^2}{2s}} - 1 \right)^s = G(x) e^{-\frac{t^2}{2}}. \end{aligned}$$

■

5 Aplicação – O Modelo Ψ -INAR(1)

Nesta secção aplicamos os resultados das secções anteriores ao modelo fortemente estacionário Ψ -INAR(1), introduzido na literatura de séries temporais inteiras por Aly e Bouzar [1]. Consideremos uma função geradora de probabilidades (f.g.p.) Ψ_t , $t > 0$, que verifica

$$\Psi_{t_1+t_2}(z) = \Psi_{t_1}(\Psi_{t_2}(z)) \text{ e } \Psi_t(0) \neq 0, \quad (6)$$

para quaisquer reais positivos t, t_1 e t_2 . Dada uma v.a. inteira positiva X , $\eta \in (0,1)$ e $\Psi_{-\ln \eta}$ a satisfazer (6) com $t = -\ln \eta$, Aly

e Bouzar [1] introduzem a v.a. operada $\eta \odot_{\Psi} X$, cuja distribuição é caracterizada, em termos da f.g.p condicionada, como

$$E(z^{\eta \odot_{\Psi} X} | X) := (\Psi_t(z))^X.$$

Mais, $E(\eta \odot_{\Psi} X) = \eta^{\delta_{\Psi}} E(X)$, onde δ_{Ψ} é uma constante. Concretamente, considerando uma sucessão de v.a.'s i.i.d. $\{Y_n\}$, independente de X e com f.g.p. Ψ_t , as v.a.'s $\eta \odot_{\Psi} X$ e $\sum_{i=1}^X Y_i$ são identicamente distribuídas.

Neste trabalho consideramos uma sub-classe da classe de Anderson, constituída pelas f.d's que verificam

$$1 - F(n) \sim n^{\xi} r^{-n} L(n), \quad n \rightarrow +\infty,$$

onde $\xi > 0$, $r > 1$ e L é uma função de variação lenta. Esta classe é aqui denotada por $\mathcal{C}_{\mathcal{A}}^*(r)$.

Consideremos o modelo Ψ -INAR(1) definido por

$$X_n = \eta \odot_{\Psi} X_{n-1} + Z_n, \quad n \geq 1, \quad (7)$$

onde $0 < \eta < 1$, $\{Z_n\}$ tem f.d. $F_{Z_n} \in \mathcal{C}_{\mathcal{A}}^*(r)$ e é independente da sucessão $\{Y_n\}$.

Teorema 5.1 *Para o modelo Ψ -INAR(1) proposto em (7) temos*

$$\lim_{n \rightarrow +\infty} P \left(\frac{S_{k_n} - k_n \mu}{\sigma \sqrt{k_n}} \leq x, M_{k_n} - d_n \leq y \right) = F_{N(0,1)}(x) \exp(-r^{-[y]}),$$

com $\mu = E(X_n)$ e $\sigma^2 = V(X_n)$.

Dem.: Dias e Temido [5] provaram que $\{X_n\}$ é uma sucessão estacionária que satisfaz as condições $D_{k_n}(u_n)$ e $D'_{k_n}(u_n)$ e que se $F_{Z_n} \in \mathcal{C}_{\mathcal{A}}^*(r)$ então também $F_{X_n} \in \mathcal{C}_{\mathcal{A}}^*(r)$. Consequentemente, devido ao que foi exposto na Secção 3, existe uma sucessão real $\{d_n\}$ tal que $F_{X_n}^{k_n}(x + d_n) \rightarrow \exp(-r^{-[x]})$, $n \rightarrow +\infty$. No caso particular da classe $\mathcal{C}_{\mathcal{A}}^*(r)$ podemos considerar $d_n = n$ e $k_n = [n^{-\xi} r^n / L(n)]$.

Assim, como a sucessão $\{X_n\}$ tem índice extremal igual a 1, concluímos que $P(M_{k_n} - d_n \leq x) \longrightarrow \exp(-r^{-[x]})$, $n \longrightarrow +\infty$.

Provemos agora que a sucessão estacionária $\{X_n\}$ verifica a condição de mistura forte com coeficiente de mistura

$$\alpha(n) = O(\varrho^{\sqrt{n}} + n^{1-d}), \quad (8)$$

onde $\varrho \in]0,1[$ e $d > 1$, seguindo de perto McCormick e Park [14]. Com efeito, para qualquer $A \subseteq \mathbb{N}_0^n$ e $B \in \mathcal{B}(\mathbb{N}_0^n)$, tem-se

$$\begin{aligned} & |P((X_1, \dots, X_n) \in A, (X_{n+2l}, X_{n+2l+1}, \dots) \in B) - \\ & - P((X_1, \dots, X_n) \in A) P((X_{n+2l}, X_{n+2l+1}, \dots) \in B)| \\ & \leq \sum_{k=n+1}^{n+l} |P((X_{n+2l-k}, X_{n+2l-k+1}, \dots) \in B | X_0 = 0) - \\ & - P((X_1, X_2, \dots) \in B)| + 2P(\overline{E}), \end{aligned} \quad (9)$$

onde $E = \{X_k = 0, \text{ para algum } k = n, \dots, n+l\}$. Provemos agora que

$$\begin{aligned} & |P((X_n, X_{n+1}, \dots) \in B | X_0 = 0) - P((X_1, X_2, \dots) \in B)| \\ & \leq C_1(\eta^{\delta_\Psi n} + n^{-d}) \end{aligned} \quad (10)$$

De facto, o primeiro membro desta desigualdade não excede

$$\begin{aligned} & \sum_{v=0}^{n-1} \left| P\left(\sum_{i=0}^{n-1} \eta^i \odot_\Psi Z_{n-i} = v\right) - P\left(\sum_{i=0}^{n-1} \eta^i \odot_\Psi Z_{n-i} + \eta^n \odot_\Psi X_0 = v\right) \right| \\ & + 2P(X_n \geq n) \\ & \leq \sum_{v=0}^{n-1} P(\eta^n \odot_\Psi X_0 \geq 1) + 2P(X_n \geq n) \leq n\eta^{\delta_\Psi n} E(X_0) + 2\frac{E(X_n^d)}{n^d}, \end{aligned}$$

onde se usou a desigualdade de Markov. Fica assim provado (10). Por outro lado, o Lema 2.5 de McCormick e Park [14] garante que $P(\overline{E}) \leq 2C_2\varepsilon^{\sqrt{l}}$, onde C_2 e $\varepsilon \in]0,1[$ são constantes adequadas. Con-

sequentemente, (9) é majorado por

$$\begin{aligned} C \sum_{k=n+1}^{n+l} (\delta^{n+2l-k} + (n+2l-k)^{-d}) + 2C_2\varepsilon^{\sqrt{l}} \\ \leq C\delta^l \frac{1-\delta^{-l}}{1-\delta^{-1}} + \frac{l}{l^d} + 2C_2\varepsilon^{\sqrt{l}} \leq C_3(\varrho^{\sqrt{l}} + l^{1-d}), \end{aligned}$$

onde $\delta := \eta^{\delta_\Psi}$ e ϱ é função de δ e de ε . Fica assim estabelecido (8). A normalidade assintótica da soma fica estabelecida, de acordo com Ibragimov e Linnik [11], se provarmos que, para algum $\beta > 0$ se tem $E(\tilde{X}_n^{2+\beta}) < \infty$ e $\sum_{n=1}^{+\infty} (\alpha(n))^{\frac{\beta}{\beta+2}} < \infty$, onde $\alpha(n)$ representa o coeficiente de mistura. Para este modelo temos $E(\tilde{X}_n^{2+\beta})$ finito porque a classe $\mathcal{C}_{\mathcal{A}}^*(r)$ admite todos os momentos finitos. Por outro lado, a série numérica de termo geral $(\varrho^{\sqrt{n}} + n^{1-d})^{\frac{\beta}{\beta+2}}$ é convergente, para $d > 1$ e $\beta > 0$ tais que $\frac{(d-1)\beta}{\beta+2} > 1$ (por exemplo, $\beta = 2$ e $d > 3$). Temos assim um Limite Central para S_n e obviamente para S_{k_n} . Seguindo novamente McCormick e Sun [15], prova-se que $\{X_n\}$ satisfaz a condição $D'_{k_n}(\sqrt{k_n}, u_n)$ desde que o coeficiente de mistura verifique $k_n\alpha(k_n^{1/6}) = o_n(1)$, $n \rightarrow +\infty$. Ora $k_n\alpha(k_n^{1/6}) = C_3(k_n\varrho^{k_n^{1/12}} + k_n^{1+\frac{1-d}{6}}) = o_n(1)$, $n \rightarrow +\infty$, para $d \geq 8$. ■

Realçamos o facto de que, se $\Psi_t(z) = 1 - e^{-t} + e^{-t}z$, então \odot_Ψ representa o operador aleatório Binomial, aqui denotado por \star . Neste caso, o modelo Ψ -INAR(1) dá lugar ao modelo estacionário INAR(1), proposto por McKenzie [16] e definido por $X_n = \eta \star X_{n-1} + Z_n$, $n \geq 1$, onde $0 < \eta < 1$. Considerando, por exemplo, $X_0 \sim NB(\lambda, p)$, com $\lambda \in \mathbb{N}$ e $0 < p < 1$, a sucessão $\{Z_n\}$ tem f.g.p. dada por $\Psi_\eta(s) = (1 - p(1 - \eta + \eta s))^\lambda (1 - ps)^{-\lambda}$. Uma vez que em McCormick e Park [14] se prova que

$$1 - F(x+n-1) \sim C(x+n)^{\lambda-1} p^{x+n}, n \rightarrow +\infty,$$

considerando $k_n = [Cp^{-n}n^{1-\lambda}]$, $\mu = \frac{p\lambda}{1-p}$ e $\sigma = \frac{\sqrt{p\lambda}}{1-p}$, tem-se

$$\lim_{n \rightarrow +\infty} P\left(\frac{S_{k_n} - k_n\mu}{\sigma\sqrt{k_n}} \leq x, M_{k_n} - n + 1 \leq y\right) = F_{N(0,1)}(x)e^{-p^{[y]}}.$$

Agradecimentos

O trabalho da primeira autora foi parcialmente financiado pela FCT - Fundação para a Ciência e a Tecnologia, pelo projeto UID/Multi/04621/2013. O trabalho da segunda autora foi parcialmente apoiado pelo Centro de Matemática da Universidade de Coimbra - UID/MAT/00324/2013, financiado pelo Governo Português através da FCT e co-financiado pelo Fundo Europeu de Desenvolvimento Regional através do Acordo de Parceria PT2020.

Referências

- [1] Aly, E. A., Bouzar, N. (2005). On a class of \mathbb{Z}_+ -valued autoregressive moving average (ARMA) processes. *REVSTAT- Statistical Journal* 6, 101–121.
- [2] Anderson, C. W., Turkman, K. F. (1991). The joint limiting distribution of sums and maxima of stationary sequences. *J. Appl. Probab.* 28, 33–44.
- [3] Anderson, C. W. (1970). Extreme value theory for a class of discrete distribution with applications to some stochastic processes. *J. Appl. Probab.* 7, 99–113.
- [4] Chow, T. L., Teugels, J. L. (1978). The sum and the maximum of i.i.d. random variables. *Proceeding of the Second Prague Symposium on Asymptotic Statistics* (Hadrec Králové), 81–92.
- [5] Dias, S., Temido, M. G. (2018) On the maxima of integer models based on a new thinning operator. *Recent Studies on Risk Analysis and Statistical Modeling*. Contributions to Statistics, Springer.
- [6] Grinevich, I. V. (1992). Max-semistable laws corresponding to linear and power normalizations. *Theory Probab. Appl.* 37, 720–721.

- [7] Grinevich, I. V. (1993). Domains of attraction of the max-semistable laws under linear and power normalizations. *Theory Probab. Appl.* 38, 640-650.
- [8] Grinevich, I. V., Khokhlov, Yu.S. (1996). The domains of attraction of semistable laws. *Theory Probab. Appl.* 40, 361-366.
- [9] Hall, A. (1998). *Extremos de sucessões de contagem*. Tese de Douto-ramento. Faculdade de Ciências da Universidade de Lisboa.
- [10] Hall, A., Temido, M. G. (2007). On the maximum term of MA and Max-AR models with margins in Anderson's class. *Theory Probab. Appl.* 51, 291-304.
- [11] Ibragimov, I. A., Linnik, Y. V. (1971). *Independent and stationary sequences of random variables*. Wolters-Noordhoff, Groningen.
- [12] Kruglov, V. M. (1972). On an extension of the class of stable distributions. *Teor. Veroyatnost. i Primenen.* 17(4), 723-732.
- [13] Leadbetter, M.R., Lindgren, G., Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag, Berlin.
- [14] McCormick, W. P., Park, Y. S. (1993). Asymptotic analysis of extremes from autoregressive negative binomial process. *J. Appl. Probab.* 29, 904-920.
- [15] McCormick, W. P., Sun J. (1993). Sums and maxima of discrete stationary processes. *J. Appl. Probab.* 40, 863-876.
- [16] McKenzie, E. (1986). Auto regressive-moving-average processes with negative binomial and geometric marginal distribution. *Advances in Applied Probability* 18, 679-705.
- [17] Temido, M.G., Canto e Castro, L. (2003). Max-semistable laws in extremes of stationary random sequences. *Theory Probab. Appl.* 47, 365-374.
- [18] Temido, M. G. (2002). Domínios de atracção de funções de distribuições discretas. In Carvalho, L., Brilhante, F., Rosado, F. (eds.): *Novos Rumos em Estatística*, 415-426, Edições SPE, Lisboa.

Estatísticas ordinais de uma amostra aleatória: O caso de dimensão de amostra com distribuição binomial negativa

Sandra Mendonça

Universidade da Madeira, Funchal, Centro de Estatística e Aplicações da Universidade de Lisboa, Lisboa, Portugal, *smfm@uma.pt*

Délia Gouveia-Reis

Universidade da Madeira, Funchal, Centro de Estatística e Aplicações da Universidade de Lisboa, Lisboa, Portugal, *delia@uma.pt*

Palavras-chave: Amostras com dimensão aleatória; Distribuição binomial negativa; Estatísticas ordinais; Divisibilidade infinita para máximos aleatórios.

Resumo: Dos muitos trabalhos sobre a distribuição do máximo de amostras com dimensão aleatória que se seguiram ao trabalho de Rachev e Resnick [3] de 1991, salientamos o trabalho de Satheesh e Sandhya [7], que generaliza alguns resultados apresentados em [3]. Dando continuidade a [1], neste trabalho são estudadas com detalhe as estatísticas ordinais de uma amostra que tem um número aleatório de elementos, tendo este número uma distribuição Binomial Negativa (r, p) . São também explorados alguns resultados limite ($p \rightarrow 0$), a max-divisibilidade infinita e a max-estabilidade, no caso em que $r = 2$.

1 Introdução

Os extremos de amostras com dimensão aleatória surgem nas mais variadas áreas, como a meteorologia, a hidrologia, a geologia, a geofísica, a eletrónica, a modelação de comportamentos na bolsa de valores, entre outras.

Consideremos $\mathbb{N} = \{1, 2, \dots\}$, $\{X_i, i \in \mathbb{N}\}$ uma sucessão de variáveis aleatórias (v.a.'s) independentes e identicamente distribuídas (i.i.d), com função de distribuição (f.d.) comum dada por F_X , $p \in (0, 1)$, N_p uma v.a. inteira e não negativa, independente de X_i , $\forall i \in \mathbb{N}$, e a k -ésima estatística ordinal descendente ($k \in \mathbb{N}$) da amostra aleatória X_1, \dots, X_{N_p} , quando a dimensão da amostra é superior ou igual a k , ($N_p \geq k$), i.e., $X_{N_p-k+1:N_p|N_p \geq k}$. Por comodidade de escrita denotaremos esta variável por $M(k, p)$:

$$M(k, p) = X_{N_p-k+1:N_p|N_p \geq k}. \quad (1)$$

Não é difícil mostrar que (cf., e.g., [1]), para x real,

$$\begin{aligned} F_{M(k,p)}(x) &= \sum_{n=k}^{\infty} \sum_{i=0}^{k-1} \binom{n}{i} [1 - F_X(x)]^i [F_X(x)]^{n-i} \frac{P(N_p=n)}{P(N_p \geq k)} \quad (2) \\ &= 1 - \sum_{i=k}^{\infty} \sum_{j=0}^{\infty} \frac{P(N_p=j+i)}{P(N_p \geq k)} \binom{j+i}{i} \overline{F_X}^i(x) F_X^j(x), \quad (3) \end{aligned}$$

com $\overline{F_X} = 1 - F_X$. Quando $k = 1$ obtém-se (de (2))

$$F_{M(1,p)}(x) = \sum_{n=1}^{\infty} F_X^n(x) \frac{P(N_p=n)}{P(N_p \geq k)} = \frac{\psi_{N_p}(F_X(x)) - P(N_p=0)}{1 - P(N_p=0)}, \quad (4)$$

onde $\psi_{N_p}(s) = E(s^{N_p})$ é a função geradora de probabilidades (f.g.p.) de N_p . A expressão (4) é facilmente generalizável para a distribuição de máximos de vetores de dimensão $d \in \mathbb{N}$, onde o *vetor máximo*, $\mathbf{M}(\mathbf{1}, \mathbf{p})$, é o vetor cujas componentes são os máximos de cada componente dos vetores envolvidos. Neste caso, tomando $\mathbf{x} = (x_1, \dots, x_d)$, $\mathbf{X} = (X_1, \dots, X_d)$, $\mathbf{M}(\mathbf{1}, \mathbf{p}) = (M(1, p)_1, \dots, M(1, p)_d)$, temos

$$F_{\mathbf{M}(\mathbf{1}, \mathbf{p})}(\mathbf{x}) = \frac{\psi_{N(s)}(F_{\mathbf{X}}(\mathbf{x})) - P(N_p = 0)}{1 - P(N_p = 0)}. \quad (5)$$

2 Quando a dimensão da amostra tem distribuição binomial negativa

Consideremos N_p uma v.a. com distribuição binomial negativa com parâmetros $r \geq 1$, inteiro, e $p \in (0,1)$, i.e., $N_p \sim \text{BN}(r,p)$, com função massa de probabilidade dada por, para $i \geq 0$, inteiro,

$$P(N_p = i) = \binom{i+r-1}{i} p^i (1-p)^r. \quad (6)$$

Neste caso, da expressão (3) resulta que (cf. [1])

$$F_{M(k,p)}(x) = 1 - \frac{P(N_{p^*} \geq k)}{P(N_p \geq k)} \quad (7)$$

com

$$p^* = \frac{p[1 - F_X(x)]}{1 - pF_X(x)}. \quad (8)$$

Notemos que, se substituirmos (8) em (6) obtemos, para $i \geq 0$, inteiro,

$$P(N_{p^*} = i) = \binom{i+r-1}{i} \left(\frac{p[1 - F_X(x)]}{1 - pF_X(x)} \right)^i \left(\frac{1-p}{1 - pF_X(x)} \right)^r, \quad (9)$$

o que conduz a (substituindo (9) em (7)),

$$F_{M(k,p)}(x) = 1 - \frac{1}{P(N_p \geq k)} \left(\frac{1-p}{1 - pF_X(x)} \right)^r \sum_{i=k}^{\infty} \binom{i+r-1}{i} \left(\frac{p[1 - F_X(x)]}{1 - pF_X(x)} \right)^i,$$

que pode ser reescrito usando funções hipergeométricas já que, quando $|g(x)| < 1$,

$$\sum_{i=k}^{\infty} \binom{i+r-1}{i} g^i(x) = g^k(x) \binom{r-1+k}{k} {}_2F_1(1, r-1+k+1; k+1; g(x)).$$

O caso $r = 1$ corresponde a ter como dimensão da amostra uma v.a. geométrica:

$$P(N_p = i) = p^i (1-p), i = 0, 1, \dots$$

A estabilidade e a divisibilidade para o caso $k = 1$, considerando a v.a. geométrica com distribuição transladada, de forma a ter como suporte \mathbb{N} , foi detalhadamente estudada por Rachev e Resnick (cf. [3]), que identificaram a variável $M(1,p)$ como sendo apropriada para descrever acontecimentos extremos que ocorrem até ao instante de uma catástrofe.

Quando $r = 2$ temos (cf. (6)), para $i = 0, 1, \dots$,

$$P(N_p = i) = (i+1)p^i(1-p)^2,$$

$$P(N_p \geq k) = \sum_{i=k}^{+\infty} (i+1)p^i(1-p)^2 = (k+1-kp)p^k \quad (10)$$

e, de (8),

$$P(N_{p^*} \geq k) = \left[k+1 - k \frac{p[1-F_X(x)]}{1-pF_X(x)} \right] \left(\frac{p[1-F_X(x)]}{1-pF_X(x)} \right)^k. \quad (11)$$

Assim, substituindo (10) e (11) em (7), obtemos

$$F_{M(k,p)}(x) = 1 - \frac{\left[k+1 - kp \frac{[1-F_X(x)]}{1-pF_X(x)} \right] \left(\frac{1-F_X(x)}{1-pF_X(x)} \right)^k}{k+1-kp}. \quad (12)$$

Na figura 1 estão representadas algumas funções de distribuição de estatísticas ordinais de ordem $N_p - k + 1$ de amostras de dimensão aleatória N_p , condicionadas à dimensão ser superior a k , nos casos em que X tem distribuição Normal(0,1) e em que X tem distribuição Exponencial(1).

Sendo X uma v.a. com função densidade de probabilidade f_X , a v.a. $M(k,p)$ terá função densidade de probabilidade dada por

$$f_{M(k,p)}(x) = \frac{(1-p)^2(k+1)k}{k+1-kp} \frac{[1-F_X(x)]^{k-1}}{[1-pF_X(x)]^{k+2}} f_X(x). \quad (13)$$

Na figura 2 estão representadas algumas funções de densidade de probabilidade de estatísticas ordinais de ordem $N_p - k + 1$ de amostras

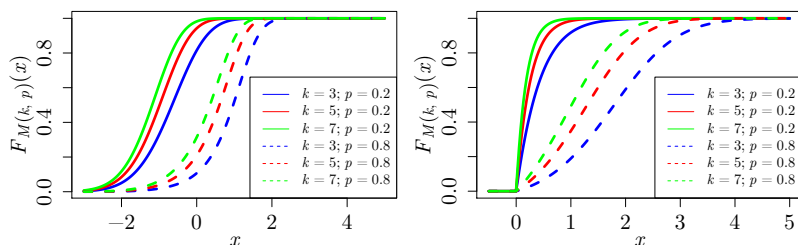


Figura 1: Algumas funções de distribuição de $F_M(k,p)$ quando X tem distribuição Normal(0,1) (direita) e quando X tem distribuição Exponencial(1) (esquerda).

de dimensão aleatória, condicionadas à dimensão ser superior a k , nos casos em que X tem distribuição Normal(0,1) e em que X tem distribuição Exponencial(1).

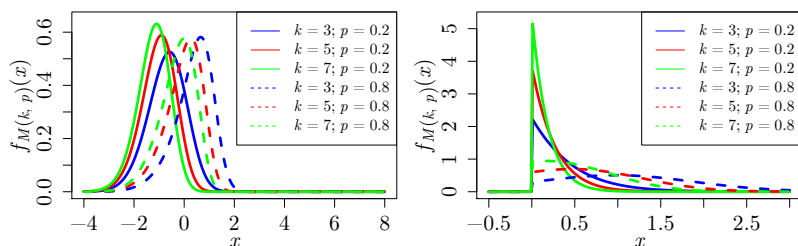


Figura 2: Algumas funções de densidade de probabilidade de $M(k,p)$ quando X tem distribuição Normal(0,1) (esquerda) e quando X tem distribuição Exponencial(1) (direita).

Se $k = 1$, obtemos o máximo da amostra, condicionado a que a amostra tenha pelo menos um elemento $M(1,p) = X_{N_p:N_p|N_p \geq 1}$ e,

de (12) e de (13), sabemos que

$$F_{M(1,p)}(x) = 1 - \frac{1}{2-p} \left[2 - \frac{p[1 - F_X(x)]}{1 - pF_X(x)} \right] \frac{1 - F_X(x)}{1 - pF_X(x)}$$

e

$$f_{M(1,p)}(x) = \frac{2(1-p)^2}{2-p} \frac{1}{[1 - pF_X(x)]^3} f_X(x).$$

Nas figuras 3 e 4 estão representadas algumas funções de distribuição (esquerda) e de funções de densidade de probabilidade (direita) de $M(1,p)$, nos casos em que X tem distribuição Normal(0,1) (Figura 3) e em que X tem distribuição Exponencial(1) (Figura 4).

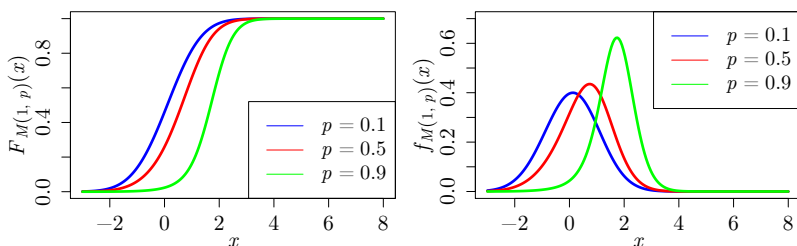


Figura 3: Algumas funções de distribuição (esquerda) e de densidade de probabilidade (direita) de $M(1,p)$ quando X tem distribuição Normal(0,1).

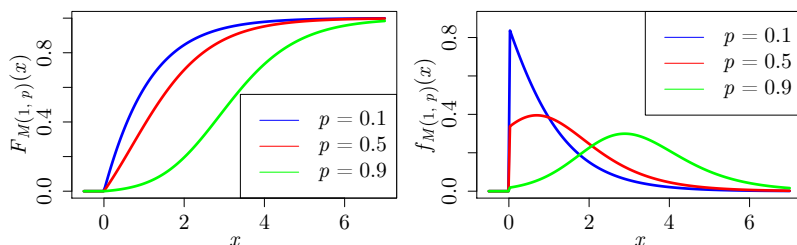


Figura 4: Algumas funções de distribuição (esquerda) e de densidade de probabilidade (direita) de $M(1, p)$ quando X tem distribuição Exponencial(1).

3 Divisibilidade infinita do máximo aleatório

Rachev e Resnick (cf. [3]) definiram um vetor aleatório Y , tomando valores em \mathbb{R}^d , como sendo infinitamente divisível para “máximos geométricos” se, para qualquer $p \in (0, 1)$, $Y \stackrel{d}{=} Y_{N_p:N_p}$, onde N_p é uma v.a. distribuída geometricamente e $\{Y_{p,j}, j \in \mathbb{N}\}$ é uma sucessão de vetores aleatórios i.i.d. e independentes de N_p .

De forma análoga, dizemos que uma v.a. Y , com função de distribuição F_Y , é uma v.a. infinitamente divisível para máximos aleatórios (i.d.p.m.a.) relativamente a uma família de distribuições de contagem N_p , $p \in (0, 1)$ se, para qualquer $p \in (0, 1)$, se tem

$$Y \stackrel{d}{=} M(1, p) = Y_{N_p:N_p | N_p \geq 1}, \quad (14)$$

onde $\{Y_{p,i}, i \in \mathbb{N}\}$ é uma sucessão de vetores aleatórios i.i.d. e independentes de N_p , com f.d. conjunta comum dada por F_p . Recorde-se a expressão obtida em (5). Sendo ψ_{N_p} uma função estritamente crescente para $s \in [0, 1]$ (com $\psi_{N_p}(0) = 0$ e $\psi_{N_p}(1) = 1$), podemos falar na função inversa de ψ_{N_p} , $\psi_{N_p}^{-1}$, e determinar a solução de (14)

tomando para distribuição de X_p

$$F_{Y_p}(x) = \psi_{N_p}^{-1}(F_Y(x)[1 - P(N_p = 0)] + P(N_p = 0)). \quad (15)$$

Notemos que, no caso em que N_p é uma v.a. positiva, temos simplesmente

$$F_{M(1,p)}(x) = \psi_{N(s)}(F_Y(x)) \text{ e } F_{Y_p}(x) = \psi_{N_p}^{-1}(F_Y(x)).$$

A f.g.p. de N_p definida por (6) é $\psi_N(s) = \left(\frac{1-p}{1-sp}\right)^r$, desde que $s < \frac{1}{p}$. A função inversa de ψ_{N_p} é

$$\psi_{N_p}^{-1}(y) = \frac{1}{p} - \frac{1-p}{p} \frac{1}{y^{1/r}} \quad (16)$$

e $P(N_p = 0) = (1-p)^r$, de onde resulta que o vetor aleatório Y é i.d.p.m.a. relativamente a $N_p \curvearrowright \text{BN}(r,p)$ se (cf. (15) e (16)) a função definida por

$$F_{Y_p}(\mathbf{x}) = \frac{1}{p} - \frac{1-p}{p} \frac{1}{[(F_Y(\mathbf{x})\{1 - (1-p)^r\} + (1-p)^r)]^{1/r}} \quad (17)$$

for uma função de distribuição. Quando $d = 1$, a expressão (17) define sempre uma função de distribuição. Quando $r = 2$ e $d = 2$ temos

$$F_{Y_p}(x,y) = \frac{1}{p} - \frac{1}{p} \frac{1}{\sqrt{\frac{p(2-p)}{(1-p)^2} F_Y(x,y) + 1}}. \quad (18)$$

Consideremos $0 \leq y_1 \leq y_2$, $0 \leq x_1 \leq x_2$ e $a_{ij} = \frac{p(2-p)}{(1-p)^2} F_Y(x_i, y_j) + 1$. Temos que:

Proposição 3.1 *A expressão (18) define uma f.d. quando*

$$\frac{1}{\sqrt{a_{22}}} + \frac{1}{\sqrt{a_{11}}} \leq \frac{1}{\sqrt{a_{21}}} + \frac{1}{\sqrt{a_{12}}},$$

sendo esta uma condição necessária, no caso em que a distribuição F_Y é simétrica e que $a_{21}^2 \leq a_{11}a_{22}$.

4 Sobre a estabilidade para máximos aleatórios

Consideremos uma v.a. X com função de distribuição F_X , $p \in (0,1)$ e N_p uma v.a. discreta tomando valores inteiros positivos, $i = 1,2,\dots$ com probabilidade $p_i = P(N_p = i)$. Consideremos ainda X_1, X_2, \dots , v.a.'s independentes e identicamente distribuídas a X e independentes de N_p . Dizemos que a distribuição F é max-estável em relação a N_p , $p \in (0,1)$, se, existirem números reais $a(p)$ e $b(p) > 0$, tais que

$$X_{N_p:N_p} \stackrel{d}{=} b(p)X + a(p), \quad (19)$$

ou, equivalentemente,

$$\psi_{N_p}(F(x)) = \sum_{i=1}^{\infty} F_X^i(x) P(N_p = i) = F_X\left(\frac{x - a(p)}{b(p)}\right). \quad (20)$$

No caso em que N_p tem um átomo em zero, como é o caso quando $N_p \sim BN(2, p)$, substituímos a v.a. $X_{N_p:N_p}$ por $M(1, k)$ (cf. (1)) e, considerando (4), reescrevemos a expressão (20) na forma

$$\frac{\psi_{N(p)}(F_X(x)) - P(N = 0)}{1 - P(N = 0)} = F_X\left(\frac{x - \beta(p)}{\alpha(p)}\right),$$

de onde resulta que F é max-estável para $N \sim BN(2, p)$ se, para algum $\alpha(p)$ e $\beta(p)$, se tem:

$$1 - \frac{1}{2-p} \left(2 - \frac{p[1 - F_X(x)]}{1 - pF_X(x)}\right) \frac{1 - F_X(x)}{1 - pF_X(x)} = F_X\left(\frac{x - \beta(p)}{\alpha(p)}\right).$$

Notemos que a estabilidade para mínimos é definida de forma semelhante – a distribuição F é min-estável em relação a N_p , $p \in (0,1)$, se existirem números reais $a(p)$ e $b(p) > 0$, tais que

$$\min(X_1, \dots, X_{N_p}) \stackrel{d}{=} b(p)X + a(p),$$

– e que a generalização para vetores aleatórios é possível, bastando para tal considerar os máximos componente a componente.

Entre os resultados já conhecidos destacamos os trabalhos de Voorn [9], Pillai [2] e de Satheesh e Nair [6].

Voorn mostrou em 1987 (*cf.* [9], Teorema 4.3) que, se a distribuição F for não degenerada e simétrica e se $\{\pi_{n,i}(p), i \in \mathbb{N}\}_{n \in \mathbb{N}}$ for uma sucessão de distribuições discretas definidas em \mathbb{N} , tais que $\lim_{n \rightarrow +\infty} \pi_{n,1}(p) = 1$, então a distribuição F é max-estável em relação a $\{\pi_{n,i}(p), i \in \mathbb{N}\}$ se, e só se, as distribuições discretas $\{\pi_{n,i}(p), i \in \mathbb{N}\}_{n \in \mathbb{N}}$ têm distribuição geométrica e se F tem distribuição logística. Voorn mostrou também que no teorema anterior poderia abdicar da simetria, juntando a estabilidade para mínimos, mostrando que (*cf.* [9], Teorema 4.4) se a distribuição F for não degenerada, se $\{\pi_{n,i}(p), i \in \mathbb{N}\}_{n \in \mathbb{N}}$ e $\{\rho_{n,i}(p), i \in \mathbb{N}\}_{n \in \mathbb{N}}$ forem sucessões de distribuições discretas definidas em \mathbb{N} , tais que $\lim_{n \rightarrow +\infty} \pi_{n,1}(p) = \lim_{n \rightarrow +\infty} \rho_{n,1}(p) = 1$, então a distribuição F é max-estável em relação a $\{\pi_{n,i}(p), i \in \mathbb{N}\}_{n \in \mathbb{N}}$, e min-estável em relação a $\{\rho_{n,i}(p), i \in \mathbb{N}\}_{n \in \mathbb{N}}$, se, e só se, as distribuições discretas $\{\pi_{n,i}(p), i \in \mathbb{N}\}_{n \in \mathbb{N}}$ são geométricas e se F tem distribuição logística, loglogística ou loglogística regressiva, colocando em destaque as mesmas distribuições encontradas por Rachev e Resnick (*cf.*, [3]). Pillai em 1991 (*cf.* [2]) mostrou que se F é uma distribuição com suporte $(0, \infty)$ é estável para máximos com respeito à distribuição Geométrica(p) ($p \in (0, 1)$) se, e só se, F tem distribuição semi-Pareto, i.e., para $x > 0$, $F(x) = 1 - \frac{1}{1+\phi(x)}$, com

$$\phi(x) = \frac{1}{p} \phi\left(p^{1/\alpha} x\right), \alpha > 0. \quad (21)$$

Um outro exemplo é dado por Sreehari em 1995 (*cf.* [8]) que generaliza a condição de estabilidade, substituindo a condição (20) por

$$\int_1^\infty F_X^t(x) dG_p(t) = F_X\left(\frac{x - a(p)}{b(p)}\right),$$

onde G_p é a função distribuição de N_p , e conclui, entre outros resultados, que as distribuições L -max-estáveis são apenas max-estáveis

com respeito a distribuições G degeneradas.

Mais recentemente, Satheesh em 2001 (*cf.* [5]) e Satheesh et. al. em 2003 (*cf.* [6]), apresentaram algumas caracterizações de variáveis $X > 0$ usando a estabilidade para máximos aleatórios. Por exemplo, mostraram que se X tem distribuição Exponencial(1), X é max-estável em relação a N_p se, e só se, N_p segue a distribuição de Sibuya(p), com função geradora de probabilidades dada por

$$\psi_{N_p}(s) = 1 - (1 - s)^p, \quad 0 < p < 1,$$

obtendo-se neste caso $\psi_{N_p}(F(x)) = F_X(px)$. Mais geralmente, $X > 0$ é max-estável em relação a N_p com a distribuição Sibuya(p) se, e só se, X tem distribuição semi-Weibull(p, α) dada por, para $x > 0$,

$$F(x) = 1 - \exp(-\phi(x)),$$

onde ϕ satisfaz a condição funcional (21), obtendo-se neste caso $\psi(F(x)) = F_X(p^{1/\alpha}x)$. Um outro caso particular provado pelos referidos autores é o caso em que $F(x) = (1 + x^{-\alpha})^{-1/k}$, com k inteiro positivo, e $\alpha > 0$. Neste caso F é estável com respeito a N_p , $p \in (0, 1)$, se, e só se, N_p tiver a distribuição de Harris($1/p, k$) (*cf.*, *e.g.*, [4]).

Agradecimentos

Investigação parcialmente suportada por fundos nacionais através da FCT - Fundação para a Ciência e a Tecnologia, Portugal, Projeto UID/MAT/00006/2013.

Referências

- [1] Mendonça, S., Pestana, D., Gomes, M.I. (2015). Randomly Stopped k th Order Statistics. In Kitsos, C.P., Oliveira, T.A., Rigas, A., Gulati, S. (eds.): *Theory and Practice of Risk Assessment: ICRA 5, Tomar, Portugal, 2013*, 249–266, Springer.

- [2] Pillai, R.N. (1991). Semi-Pareto Processes, *J. Appl. Prob.* 28, 461-465.
- [3] Rachev, S.T., Resnick, S. (1991). Max-geometric infinite divisibility and stability. *Commun. Stat.-Stoch. Models* 7, 191–218.
- [4] Sandhya, E., Sherly, S., Raju, N. (2005). Harris Family of Discrete Distributions, arXiv:math/0506220.
- [5] Satheesh, S. (2001). Stability of random sums and extremes, PhD. Thesis, Cochin University of Science and Technology (consultado em: <https://dyuthi.cusat.ac.in/xmlui/handle/purl/776>, 25/07/2017).
- [6] Satheesh, S., Nair N.U. (2002). A Note on Maximum and Minimum Stability of Certain Distributions, *Calcutta Statistical Association Bulletin* 53, 249–252
- [7] Satheesh, S., Sandhya, E. (2014). N-max infinite divisibility and N-max stability. arXiv:1405.4782 [math.PR].
- [8] Sreehari, M. (1995). Maximum stability and a generalization, *Statistics & Probability Letters* 23, 339-342.
- [9] Voorn, W.J. (1987). Characterization of the logistic and loglogistic distributions by extreme value related stability with random sample size, *J. Appl. Prob.* 24, 834–851.

Autores

- Abreu**, Ana Maria, 145
Afonso, Anabela, 75
Alves, Paulo Marques, 235
Botelho, Maria do Carmo, 235
Cabral, Ivanilda, 129
Caeiro, Frederico, 59, 129
Canto e Castro Loura,
Luísa, 203
Cardoso, Margarida G. M. S.,
45
Cordeiro, Clara, 189
Dias, Sandra, 249
Ferreira, Ana Sousa, 13
Ferreira, Fátima, 117
Figueiredo, Mário A. T. , 45
Godinho, Ana Rita, 1
Gomes, M. Ivette, 129
Gonçalves, Esmeralda, 105
Gonçalves, Elsa, 89
Gouveia-Reis, Délia, 263
Isaac, Victória, 161
Malaca, Carlos, 203
Marques, Anabela, 13
Martins, Antero, 89
Martins, Cristina, 105
Mendes-Lopes, Nazaré, 105
Mendes, Marcos Huber, 175
Mendes, Zilda, 1
Mendonça, Sandra, 263
Neves, M. Manuela, 189
Oliveira, Manuela, 59
Pacheco, António, 117
Paulino, Carlos Daniel, 31
Pereira, Dulce G., 75
Pereira, Júlio César, 161
Ribeiro, Helena, 117
Rocha, Cristina, 1, 145
Rodrigues, Nuno, 203
Salgueiro, Maria de Fátima,
221
Sanfins, Marco Aurélio, 175
Santos, Joaquim, 203
Silva, Domingos, 59
Silva, Giovanni L., 31, 161
Silvestre, Cláudia, 45
Sousa-Ferreira, Ivo, 145
Sousa, Reinaldo Castro, 175
Temido, Maria da Graça, 249
Vicente, Paula C.R., 221

